| | |
|---|---|
| **Titre:** Title: | Nouvelle approche de maîtrise de processus intégrant les cartes de contrôle multidimensionnelles et les graphes en coordonnées parallèles |
| **Auteur:** Author: | Shaima Tilouche |
| **Date:** | 2017 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:** Citation: | Tilouche, S. (2017). Nouvelle approche de maîtrise de processus intégrant les cartes de contrôle multidimensionnelles et les graphes en coordonnées parallèles [Ph.D. thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/2953/ |

UNIVERSITÉ DE MONTRÉAL

# NOUVELLE APPROCHE DE MAÎTRISE DE PROCESSUS INTÉGRANT LES CARTES DE CONTRÔLE MULTIDIMENSIONNELLES ET LES GRAPHES EN COORDONNÉES PARALLÈLES

SHAIMA TILOUCHE
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INDUSTRIEL)
DÉCEMBRE 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

NOUVELLE APPROCHE DE MAÎTRISE DE PROCESSUS INTÉGRANT LES CARTES DE CONTRÔLE MULTIDIMENSIONNELLES ET LES GRAPHES EN COORDONNÉES PARALLÈLES

présentée par : TILOUCHE Shaima
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

M. FRAYRET Jean-Marc, Ph. D, président
M. BASSETTO Samuel-Jean, Doctorat, membre et directeur de recherche
M. PARTOVI NIA Vahid, Doctorat, membre et codirecteur de recherche
M. DANTAN Jean-Yves, Doctorat, membre
M. PILLET Maurice, Doctorat, membre externe

# DÉDICACE

*À mes parents qui m'ont inculqué le sens de la responsabilité et de la confiance en soi,*
*qui ont guidé mes pas vers la réussite et qui n'ont cessé de me combler par leur amour,*
*À mes frères, Beha et Dhia, qui m'ont soutenu au cours des moments les plus difficiles,*
*À mon mari pour son soutien, ses encouragements, son affection et son amour,*
*À la mémoire de mes grands parents qui ont été et seront toujours dans mon esprit et dans mon cœur,*
*À tous ceux qui me sont chers et que ma réussite leur tient à cœur.*
*Sans vous, je n'aurais pas réussi,*

# REMERCIEMENTS

Avant tout développement, il apparaît opportun de commencer cette thèse de doctorat par des remerciements, à ceux qui m'ont beaucoup appris, et même à ceux qui ont eu la gentillesse de faire de cette expérience un moment très profitable.

Tout d'abord, je souhaite adresser mes remerciements à mon directeur, Professeur Samuel-Jean Bassetto, pour m'avoir accordé toute sa confiance, pour tout le temps qu'il m'a consacré tout au long de cette période et pour son soutien continu scientifique, technique et moral. Je le remercie particulièrement pour ses efforts continus de créer un cadre de travail amical et agréable au CIMAR-LAB et pour sa façon innovante de voir les choses.

Je tiens, également à remercier mon codirecteur, Docteur Vahid Partovi Nia, pour son temps, ses conseils précieux et pour toutes les connaissances qu'il m'a transmises particulièrement, dans le domaine des statistiques.

Je tiens à témoigner toute ma reconnaissance, aux employés de Teledyne Dalsa, pour leur collaboration et leur soutien et particulièrement Docteur Christian Veilleux et Docteur Khaled Amir Belarbi.

Mes remerciements s'adressent, aussi, à mes collègues de CIMAR-LAB, d'avoir participé à créer une ambiance de travail assez motivante et d'avoir été à mes cotés lors des moments difficiles. Vous n'êtes pas que des collègues de bureau, vous êtes des amis, Julie Charron-Latour, Antoine Murry, Shima Saffar et Sheida Shams.

Je souhaite, également remercier les membres du jury, Professeur Jean-Marc Frayret, Professeur Jean-Yves Dantan et Professeur Maurice Pillet d'avoir accepté de réviser cette thèse.

Je n'oublie pas de mentionner que ce projet n'aurait pas été mené à terme sans la participation du Professeur Patrick desjardins, du Professeur Paul Charette et Teledyne Dalsa à son soutien financier.

Finalement, j'exprime ma gratitude à toutes les personnes qui ont participé de près ou de loin à la réalisation de ce projet de prés ou de loin.

# RÉSUMÉ

Dans une entreprise nord américaine type, les coûts de non-qualité sont en moyenne de 20% de son chiffre d'affaires. Ces coûts sont certainement élevés et ils ne peuvent pas être, totalement, éliminés. Toutefois, les entreprises peuvent les réduire grâce à une meilleure maitrise des processus manufacturiers et à un meilleur contrôle qualité. Ces taches sont primordiales pour garantir l'efficacité des processus de fabrication et pour améliorer la qualité des produits. En effet, la qualité des produits est reliée aux paramètres machines. Cependant, actuellement dans l'industrie, les paramètres machines et les variables des produits sont contrôlés séparément omettant ainsi les relations qui peuvent exister entre eux. La vérification individuelle séparée peut être longue et complexe. Elle peut mener à la non-détection de certains défauts ou encore à la génération de certaines fausses alarmes. En effet, la prise en compte des relations entre les paramètres des équipements et/ou les variables des produits est indispensable. Pour tenir compte des dépendances entre les variables et paramètres, plusieurs auteurs ont proposé des cartes de contrôle multidimensionnelles, telles que les versions multidimensionnelles des cartes connues telles que MEWMA, CUSUM et Hotelling. Ces cartes ont un problème majeur. Elles supposent que les données proviennent d'une distribution normale, ce qui n'est pas toujours le cas. D'autres versions des cartes de contrôle ne supposent pas la normalité des données, mais supposent que leur distribution est connue. Or, peu d'industriels connaissent ce genre d'informations. D'autres techniques de contrôle de processus ou de détection de défauts ont été suggérées. Ces techniques sont soient des techniques basées sur des algorithmes d'apprentissage statistique ou de data mining soient des cartes de contrôle qui ne dépendent pas de la distribution des données. Ces outils ont montré des résultats assez intéressant en termes de détection de défauts et de génération de fausses alarmes. Par contre, elles fonctionnent comme une sorte de boite noire. Si un défaut est détecté, le diagnostic doit passer par des cartes de contrôle monodimensionnelles et doit idéalement se faire par un expert. Ces outils proposent rarement un support visuel de diagnostic. Ceci peut être du au fait que les graphes multidimensionnelles sont généralement méconnus ou, parfois, difficile à interpréter. Ainsi, ils sont rarement exploités dans le développement des outils de contrôle.
Dans ce document, nous proposons d'intégrer un type de graphes multidimensionnelles, les coordonnées parallèles avec les concepts des outils de contrôle pour soutenir le contrôle qualité. Nous proposons un outil visuel de contrôle de processus, qui est ne dépend pas de la distribution des données et qui tient en compte les relations entre les variables considérées. Cet outil permet de faire le diagnostic d'un défaut détecté. Cet outil permet de générer deux types de cartes de contrôle multidimensionnelles selon la disponibilité des données historiques.

Les deux cartes sont visualisées en coordonnées parallèles. La première version est proposée pour le cas ou un nombre assez important d'observations historiques est disponible. Elle est basée sur la visualisation des limites multidimensionnelles de la zone de fonctionnement appelée best operating zone. Cette zone est encore répartie en plusieurs zones de fonctionnement. La deuxième version est adaptée au cas ou le nombre de données historiques est limité. Elle est basée sur la caractérisation de la zone de fonctionnement à l'aide des graphes de densité. Avant de caractériser les zones de fonctionnement, pour garantir une représentation optimisée des variables en coordonnées parallèles, un arrangement des variables dans l'objectif de souligner les relations entre les variables ou d'améliorer la détection des segments de fonctionnement est réalisé. Un cadre général d'arrangement de variables est proposé. Ce cadre dépend de l'objectif d'arrangement.

Pour conclure, la conception des cartes de controle passe par 3 étapes principales :

- L'arrangement des variables ;
- La caractérisation de la zone opérationnelle (zone de fonctionnement) ;
- la représentation et la classification des nouvelles observations.

Chaque étape du développement de l'outil est évalué à l'aide d'une ou plusieurs bases simulées ou réelles pour montrer les avantages et les limitations des algorithmes et des outils suggérés. L'algorithme d'arrangement des variables montre sa capacité à détecter les dépendances entre les attributs et aussi à séparer les données. Les cartes de contrôle basées sur la best operating zone (première version) offre un taux de détection de défauts assez élève (environ 76% pour la base de données de spam) et un taux de fausses alarmes acceptable comparé aux cartes d'Hotelling. De plus, ces cartes montrent une performance comparables voire meilleure que celles des cartes d'Hotelling selon le critère de la longueur opérationnelle moyenne (ARL). Les cartes de contrôle densité, développées avec un nombre de données limitées, montrent un taux de classification assez intéressants comparées aux réseaux de neurones et aux cartes d'Hotelling. Elles donnent un taux de classification correcte autour de 75% en se basant sur des cartes développées avec 100 observations historiques. Le même taux est trouvé avec les réseaux de neurones mais avec 300 observations historiques (d'apprentissage). Le taux de classification des cartes d'Hotelling est, significativement, plus faible que celui des cartes densité et des réseaux de neurones. Les tests montrent que les solutions proposées s'alignent avec les objectifs pour lesquelles elles ont été proposées, notamment pour l'aspect visualisation et diagnostic des cartes de contrôle.

# ABSTRACT

Quality control and process monitoring are very important task for manufacturing processes. They guaranty the efficiency of the manufacturing process and the quality of the final products. Final product quality is directly related to equipment parameters.Despite the dependency between the process parameters and the product variables, they are separately monitored in most of the current industries. Generally, each parameter or variable is monitored in individual process control chart which might make the control a longer and more complex. This might, also, be very misleading. It might lead to the non-detection of some faults or to the generation of false alarms. Actually, taking into account the dependencies between product variables and process parameters is necessary. In order to do so, many authors suggest multivariate versions of known process control charts such as MEWMA, CUSUM and Hotelling. These charts have a major problem, that is they are under a very restrictive assumption, as they consider that all the variables and parameters follow a normal distribution. Authors suppose that somehow the central limit theorem will solve the problem of data non-normality. This is true when the charts are proposed for monitoring statistics such as the mean or the standard deviation, but not accurate when it is about monitoring individual observations. Some authors suggest techniques that do not suppose the normality of the data but that suppose that it is known. Few industrials know this kind of information, i.e. statistical characteristics of data. As an improvement of the parametric charts, non-parametric process control tools were proposed. These tools are either techniques based on machines learning or data mining algorithms or distribution free control charts. They show interesting results in fault detection and false alarm generation. However, they work as a black box. It is difficult to understand or interpret the obtained results. If any fault is detected, the diagnosis needs to be proceeded by an expert usually supported by monodimensional charts. Actually, practitioners are still not familiar with multidimensional graphs. In this thesis, we introduce a visual distribution free multidimensional process control tool that takes into account the dependencies between the different variables and parameters. This tool integrates parallel coordinates with the concepts of process control tools. So, it enables fault detection and also diagnosis as it conceives two types of visual control charts, depending on the availability of the historical (training) data. Both charts are visualized in parallel coordinates. The first version is proposed for the case where the training dataset is large. It is based on the visualization of control limits, i.e. the limits of the best operating zone. This zone that contains all possible functional observations is, then, divided into small functional zones in a way that the probability of not detecting a fault is reduced. The second version of chart is

proposed for the case where the number of historical data is limited. The characterization of the operating zone is based on density graphs. However, before characterizing the operating zone, a variable reordering is applied to ensure an optimized representation of the variables in the parallel coordinate graph. The objective of this step is to highlight relations among variables, highlight data structure and help cluster detection. A general variable reordering framework is presented. It depends on the objective of the reordering.

To conclude, conceiving a control chart, as it is proposed in this thesis goes through 3 steps:

— variable reordering;

— characterizing the functional (operating) zone;

— representing and classifying the new observations.

Each step of the development of the tool is evaluated based on different databases to analyze the advantages and limitations of the proposed algorithms.

The suggested variable reordering framework shows its capacity to adapt to the objective of reordering. Twos objective were studied, highlighting variable dependence and data separation.

The results obtained for the first version of the control chart are comparable (or better) than Hotelling chart, 76% of correct classification compared to 69% for Hotelling charts (for SPAM data). This is confirmed when the average run lengths are compared (ARL). Moreover, the density charts give, also, interesting results compared to Hotelling charts and neural networks. It reaches 75% of correct classification rate with 100 historical observations, whereas, neural networks reach the same rate with 300 observations. Hotelling charts do not give interesting results when the number of historical observations is limited.

Besides, their good performance, the proposed charts provide a visual support that enables the interpretation of the results and also, the diagnosis of the detected faults which is not offered by the other techniques.

# TABLE DES MATIÈRES

DÉDICACE . . . . . iii

REMERCIEMENTS . . . . . iv

RÉSUMÉ . . . . . v

ABSTRACT . . . . . vii

LISTE DES TABLEAUX . . . . . xi

LISTE DES SIGLES, ABRÉVIATIONS ET NOTATIONS MATHÉMATIQUES . . . xv

CHAPITRE 1 INTRODUCTION . . . . . 1

CHAPITRE 2 REVUE DE LITTÉRATURE . . . . . 9
2.1 Outils de contrôle multidimensionnels . . . . . 9
2.1.1 Revue des outils de maîtrise des processus . . . . . 9
2.1.2 Cartes de contrôle multidimensionnelles à support visuel . . . . . 14
2.1.3 Critères de performance des cartes de contrôle . . . . . 17
2.2 Graphes en coordonnées parallèles . . . . . 18
2.2.1 Présentation des graphes en coordonnées parallèles . . . . . 18
2.2.2 Dualité $2D$ et coordonnées parallèles . . . . . 19
2.2.3 Réordonnancement des variables . . . . . 24
2.2.4 segmentation . . . . . 28
2.2.5 Graphes de densité en coordonnées parallèles. . . . . 29
2.3 Conclusion . . . . . 31

CHAPITRE 3 RÉARRANGEMENT DES VARIABLES . . . . . 32
3.1 Information générale . . . . . 32
3.2 Optimisation de l'ordre . . . . . 34
3.3 Application du cadre général pour objectif de dépendance . . . . . 35
3.4 Application du cadre général pour objectif de séparation . . . . . 36
3.5 Explication des étapes de l'algorithme d'arrangement de variables à travers un cas simple : critère de dépendance . . . . . 38
3.6 Conclusion . . . . . 39

CHAPITRE 4 CARTES DE CONTRÔLE MULTIDIMENSIONNELLES . . . . . 41

4.1 Projection de points en coordonnées parallèles dans un espace Cartésien à 2 dimensions . . . 42
4.2 Développement des cartes BOZ . . . 45
4.2.1 Identification de l'enveloppe de la Best Operating Zone . . . 45
4.3 Segmentation . . . 48
4.4 Validation croisée . . . 51
4.5 Automatisation de la détection de classification des nouvelles observations . . . 52
4.6 Évaluation de la performance de la carte de contrôle développée . . . 53
4.7 Graphes de densité en coordonnées parallèles . . . 55
4.8 Conclusion . . . 59

CHAPITRE 5 RÉSULTATS . . . 60
5.1 Évaluation de l'ordre des variables en coordonnées parallèles . . . 60
5.1.1 Critère de dépendance : Base de données de la qualité de vin . . . 60
5.1.2 Critère de séparation : Base de données génétique . . . 65
5.2 Évaluation des cartes de contrôle basées sur la BOZ . . . 68
5.2.1 Présentation et explication des cartes de contrôle : base de données simulées . . . 68
5.2.2 Application à un cas réel : base de données de SPAM . . . 77
5.3 Comparaison de la performance de la carte BOZ avec la carte d'Hotelling (ARL) 81
5.4 Conclusion . . . 81
5.5 Évaluation des cartes de Contrôle de densité . . . 82
5.5.1 Base de données . . . 83
5.5.2 Présentation de la carte de contrôle densité . . . 84
5.5.3 Description du test . . . 84
5.5.4 Évaluation des cartes de contrôle densité . . . 86
5.5.5 Lien avec les plans d'expérience . . . 88
5.5.6 Conclusion . . . 88

CHAPITRE 6 CONCLUSION . . . 90
6.1 Synthèse des travaux . . . 90
6.2 Limitations de la solution proposée . . . 91
6.3 Améliorations futures . . . 92

RÉFÉRENCES . . . 93

# LISTE DES TABLEAUX


Tableau 5.1 Relations qui relient les différents attributs. Par exemple la ligne 4, colonne 6 montre la relation qui relie $D.Cool$ et $T2 : T2 = 18D.Cool^2+34$. 69

Tableau 5.2 Taux de classification correcte des différentes cartes étudiées. . . . . 75

Tableau 5.3 Description des données de la base de SPAM. . . . . 78

Tableau 5.4 Résultats de classification des données SPAM avec les cartes BOZ et la carte d'Hotelling. . . . . 80

Tableau 5.5 Valeurs de L'ARL pour différentes covariances entre les variables et pour différentes distances de la moyenne. . . . . 82

## LISTE DES FIGURES


Figure 1.1 Figure illustrant les limites de contrôle d'une température et d'une pression d'un cycle Carnot sans et avec prise en compte des relations entre ces paramètres. . . . . 3

Figure 2.1 Figure illustrant une région opérationnelle elliptique représentée en rouge en coordonnées parallèles. . . . . 16

Figure 2.2 Figure illustrant la dualité entre un graphe à 3 dimensions en plan Cartésien et en coordonnées parallèles pour 2 points $A(1, 1, 2.5)$ et $B(-1, 0, 1.5)$. . . . . 19

Figure 2.3 Projection d'observations en coordonnées parallèles dans un plan Cartésien. . . . . 20

Figure 2.4 Figure illustrant la dualité entre un plan Cartésien à 2 dimensions et les coordonnées parallèles pour les fonctions linéaires. De gauche à droite, nous avons, une fonction linéaire pente négative, une fonction constante et une fonction à pente négative avec et sans bruit. Les figures d'en haut sont représentées dans un plan Cartésien. L'équivalent en coordonnées parallèles est représenté dans les figures en bas. . . . . 22

Figure 2.5 Figure illustrant la dualité entre $2D$ et coordonnées parallèles pour des fonctions non-linéaires. De gauche à droite sont représentées les fonctions suivantes : fonction carré à coefficient positif puis négatif, fonction cubique à coefficient positif et négatif et une fonction périodique (sinusoïdale). . . . . 23

Figure 2.6 Figure illustrant la dualité entre $2D$ et coordonnées parallèles pour la fonction $f : x \to \frac{1}{x}$ sur l'intervalle $[-1, 1]$. . . . . 24

Figure 2.7 Figure illustrant la perception des données séparables ou/ et corrélées en coordonnées Cartésiennes et en coordonnées parallèles.. . . . . 25

Figure 2.8 Exemple de données avant (figure de gauche) et après réarrangement manuel (figure de droite). Le réarrangement est réalisé dans l'objectif visualiser les dépendances entre les variables. . . . . 26

Figure 2.9 Exemple de données avant (figure de gauche) et après réarrangement manuel (figure de droite). Le réarrangement est réalisé dans l'objectif d'améliorer la séparation visuelle des données. . . . . 27

Figure 4.1 Étapes de développement des cartes de contrôle. . . . . 41

Figure 4.2 Données avant et après standardisation. . . . . 43

Figure 4.3 Figure expliquant la projection de points représentés en coordonnées parallèles dans un espace Cartésien à 2 dimensions. . . . . 44

| | | |
|---|---|---|
| Figure 4.4 | Projection de points en coordonnées parallèles dans un plan Cartésien. | 45 |
| Figure 4.5 | Étapes de conception des cartes BOZ. . . . . | 45 |
| Figure 4.6 | Exemple de formes qui apparaissent en coordonnées parallèles. . . . . | 46 |
| Figure 4.7 | Exemple montrant les limites de la BOZ. . . . . | 47 |
| Figure 4.8 | Exemple d'une zone intérieure vide qui ne fait pas partie de la BOZ. . | 48 |
| Figure 4.9 | Étapes d'une réplication de validation croisée. . . . . | 51 |
| Figure 4.10 | Étapes de conception des cartes BOZ. . . . . | 55 |
| Figure 4.11 | Données multidimensionnelles projetées dans un plan Cartésien bidimensionnel. . . . . | 57 |
| Figure 5.1 | Figure montrant les données de vins blancs réordonnées avec la statistique d'information mutuelle. Les valeurs en bleu représentent les valeurs de l'information générale entre 2 paires d'attributs. . . . . | 62 |
| Figure 5.2 | Figure montrant les données de vins blancs réordonnées avec la statistique d'information mutuelle. Les valeurs en bleu entre chaque couple d'attributs adjacents sont les valeurs de l'information générale $GI(x_i, x_j)$. De haut en bas et de droite à gauche sont représentées les données ordonnées avec la statistique Freeman-Tukey, Neyman, Cressie et Pearson. | 63 |
| Figure 5.3 | Figure illustrant le comportement des courbes des fonctions qui permettent d'obtenir les différentes statistiques étudiées autour de 1. . . | 64 |
| Figure 5.4 | Figure montrant les données génétiques réordonnées avec le critère de séparation (figure en haut) et avec la corrélation de Pearson (figure d'en bas). . . . . | 66 |
| Figure 5.5 | Figure illustrant les données dans l'ordre initial de simulation, les données ordonnées selon le critère de dépendance ; mesure d'information mutuelle et statistique de Pearson, et le critère de séparation. . . . . | 70 |
| Figure 5.6 | Figure illustrant les limites de BOZ. . . . . | 71 |
| Figure 5.7 | Figure représentant la première catégorie de défauts qui sont ceux qui dépassent la limite inférieure ou supérieure d'une ou de plusieurs variables. . . . . | 72 |
| Figure 5.8 | Figure illustrant le deuxième type de dérives soit les observations qui ne respectent pas les limites des relations entre les variables. . . . . | 72 |
| Figure 5.9 | Figure illustrant les dérives détectables à l'aide des segments de fonctionnement. . . . . | 73 |
| Figure 5.10 | Figure illustrant le deuxième type de dérives soient les observations qui ne respectent pas les limites des relations entre les variables. . . . . | 75 |

Figure 5.11 Figure illustrant le deuxième type de dérives soit les observations qui ne respectent pas les limites des relations entre les variables. . . . . 76
Figure 5.12 Données SPAM ordonnées avec le critère de séparation (figure en haut) et le critère de dépendance (figure de dessous). . . . . 79
Figure 5.13 Exemple de carte de contrôle BOZ avec les données SPAM. . . . . 80
Figure 5.14 L'évolution de l'ARL en fonction de la distance dérivant par rapport à la moyenne sous contrôle. De gauche à droite, $\sigma$ varie de 0.2, 0.5 et 0.8. 83
Figure 5.15 Figure illustrant les données dans l'ordre initial de simulation, les données ordonnées selon le critère de dépendance ; mesure d'information mutuelle et statistique de Pearson, et le critère de séparation. . . . . 85
Figure 5.16 Méthode suivie pour tester les cartes de contrôle densité : courbe d'apprentissage. . . . . 86
Figure 5.17 Courbe d'apprentissage des différents algorithmes testés, les cartes densité, BOZ et Hotelling et les réseaux de neurones . . . . . 87
Figure 5.18 Évolution du taux de fausses alarmes des différents algorithmes testés, les cartes densité, BOZ et Hotelling et les réseaux de neurones en fonction du nombre d'observations de la base de données historiques. 88

# LISTE DES SIGLES, ABRÉVIATIONS ET NOTATIONS MATHÉMATIQUES

| | |
|---|---|
| BOZ | Best Operating Zone ; |
| SVM | Support vector machines ; |
| ARL0 | Average Run Length, longueur opérationnelle moyenne sous contrôle ; |
| ARL1 | Average Run Length, longueur opérationnelle moyenne hors contrôle ; |
| UCL | Upper Control Limit, limite supérieure de contrôle ; |
| LCL | Lower Control Limit, limite inférieure de contrôle ; |
| $x$, $x_i$ | une variable ; |
| $\mathbf{x}$, $\mathbf{y}$ | un vecteur ; |
| $\mathbf{X}$ ou $\mathbf{Y}$ | une matrice ; |
| $X$ | une variable aléatoire ; |
| $\max\{x_i\}$ | la valeur maximale que peut prendre la variable $x_i$ ; |
| $\min\{x_i\}$ | la valeur minimale que peut prendre la variable $x_i$ ; |
| $\bar{\mathbf{x}}_h$ | vecteur de moyenne ; |
| $\mathbf{o}$ | vecteur représentant une nouvelle observation (à controler) ; |
| $T^2$ | statistique d'Hotelling ; |
| $\overline{x}_i$ | moyenne de la variable $x_i$ ; |
| $\mathbf{\Sigma}$ | matrice de variance-covariance ; |
| $\sigma_i$ | variance de la variable $x_i$ ; |
| $\sigma_i j$ | covariance entre la variable $x_i$ et $x_j$ ; |
| $\mu$ | vecteur de moyenne réelles ; |
| $F_l, k$ | distribution de Fisher de degrés de liberté $l$ et $k$ ; |
| $beta$ | distribution Beta ; |
| $F_X(.)$ | fonction de répartition de la variable $X$ ; |
| $f_X(.)$ | fonction de densité de la variable $X$ ; |
| $GI(x_1, x_2)$ | information générale des variables $x_1$ et $x_2$ ; |
| $G(.)$ | fonction univariée ; |
| $G''(.)$ | dérivée seconde de la fonction $G$ ; |
| $F$, $H$ | mesures de proabilités ; |
| $\odot$ | le produit Hadamart ; |
| $\mathbf{1}$ | vecteur unitaire ; |
| $\mathbf{a}_i^\top$ | vecteur transposé du vecteur $\mathbf{a}_i$ ; |
| $\#_C$ | nombre d'élément dans la classe $C$ ; |
| $C_i^{x_1}$ | classe $i$ de la variable $x_1$ ; |

| | |
|---|---|
| $\epsilon$ | un nombre assez petit ; |
| $f$ ou $f_i$ ou $f_l^k$ | fonction univariée ; |
| $K_H()$ | fonction définie à partir d'une fonction noyau $K$ ; |

# CHAPITRE 1 INTRODUCTION

Selon Fortin (1990), l'entreprise nord-américaine type, dans le secteur manufacturier, perdrait chaque année en moyenne, 20% de son chiffre d'affaires à cause de la non-qualité. Pour la province du Québec au Canada, ce coût serait de 15 milliards de dollars. Selon certains auteurs, ces coûts peuvent atteindre jusqu'à 30% dans certaines entreprise (Abouzahir et Gidel, 2003). Le coût de la non qualité inclut plusieurs coûts reliés à des opérations ayant lieu en interne dont des coûts de prévention, des coûts de détection de défauts et des coûts défaillance interne (rebuts, retouches, etc.). Mais, il inclut, également, des coûts externes comme les coûts des retards de livraison, les coûts de gestion litiges avec le client, les coûts logistiques d'acheminement ou de stockage du produit de remplacement et les coûts du préjudice commercial (coût de dégradation de l'image de l'entreprise ou coût de la perte de clients).

Les coûts externes notamment du retour de produits et du préjudice commercial peuvent être significativement réduits grâce à une meilleure maîtrise de la qualité. Celle-ci repose sur la maîtrise des processus de production, qui elle même repose sur une maîtrise fine des ressources de production (Bleakie et Djurdjanovic, 2013). À cause de la complexité des processus de fabrication, le diagnostic des équipements défectueux ne peut, généralement, pas être effectué immédiatement. En effet, cette action nécessite des connaissances professionnelles, en termes de matériaux et d'ingénierie (Chen *et al.*, 2011). Au niveau des équipements de production, plusieurs capteurs sont placés pour réguler et contrôler la gamme de fabrication. Ceci rend disponibles de très nombreuses données qui peuvent être employées pour caractériser le fonctionnement des équipements. Ces données sont exploitées pour assurer la maitrise de processus de production, ou pour détecter les défauts ou encore pour évaluer l'état de santé des équipements. La maîtrise des processus peut être réalisée soit par des cartes de contrôle statistiques ou par des techniques d'apprentissage statistiques ou d'exploration de données industrielles (data mining). Cependant, actuellement, les cartes de contrôle monodimensionnelles sont les outils de contrôle les plus utilisés dans l'industrie pour la maîtrise des processus. La maîtrise des processus est réalisée en se basant sur ces cartes et sur l'expérience et la connaissance des experts du domaine. L'expert analyse les cartes soit pour réagir et réguler l'équipement ou le processus afin d'éviter la dérive, soit pour arrêter l'équipement et le réparer. Ceci est le cas du partenaire industriel de ce projet de recherche, Teledyne Dalsa.
Teledyne Dalsa est un chef de file international dans la haute performance de l'imagerie numérique et des semi-conducteurs. Le C2MI est un centre internationale de collaboration et d'innovation dans le secteur des MEMS et de l'encapsulation. Il est le maillon essentiel entre

la recherche appliquée et la commercialisation de produits dans la microélectronique [1]. Le C2MI est en partenariat entre plusieurs universités dont l'école polytechnique de Montréal et les industriels dont Teledyne Dalsa.

Le processus de production chez Teledyne Dalsa comme dans les usines de semi-conducteurs est un processus complexe (Lynn, 2011). Les galettes de silicium (wafers) doivent passer par des centaines d'opérations de fabrication et de contrôle avant d'obtenir les produits finaux (Kang *et al.*, 2011). Le contrôle est réalisé par un expert du domaine soutenu par des cartes de contrôle monodimensionnelles.

Les cartes de contrôle monodimensionnelles ont été proposées par Shewhart en 1926 (Bakir, 2004) pour surveiller les paramètres des processus de fabrication. Une carte de contrôle statistique de la qualité est une procédure conçue pour surveiller la stabilité d'un processus en traçant une séquence de statistiques sur un graphique. Ce graphique utilise une ligne centrale et une ou plusieurs limites de contrôles établies statistiquement (Bakir, 2004). La statistique en question peut soit concerner les observations individuelles, par exemple, la température à l'instant $t$, soit un sous-groupe d'observations, par exemple, une moyenne ou une variance d'un sous-groupe d'observations. Malheureusement, la plupart des méthodes de maîtrise statistique des processus est basée sur la représentation d'un petit nombre de paramètres machines ou de mesures sur les produits finaux, habituellement, examinés séparément (MacGregor et Kourti, 1995). Les cartes de contrôle monodimensionnelles comme la carte $T2$ d'Hotelling, la carte CUSUM (Woodward et Goldsmith (1964)) et/ou la carte EWMA (Roberts (1959) ; Hunter (1986)) sont utilisées, en général, pour contrôler séparément les mesures clés des produits finaux définissant, ainsi la qualité du produit. Le contrôle séparé des mesures et paramètres peut empêcher la détection et la compréhension des dépendances entre les variables. Dans les processus industriels, cette approche n'est pas toujours efficace. La compréhension des relations entre les paramètres des processus et des équipements est indispensable, mais assez souvent omise (Bassetto et Siadat, 2009). Le contrôle séparé peut mener à un contrôle erroné. La figure 1.1 illustre un exemple soulignant l'importance de la prise en compte des relations entre les paramètres machines.

Soient $T$ et $P$ la température et la pression d'un cycle de Carnot ayant respectivement des limites inférieures et supérieures de 558 et 562 et de 0.22 et 0.38. En considérant séparément les limites, la zone de points sous contrôle est délimitée par les lignes (en bleu). Or, si la relation entre la température et la pression, $(T - 560)^2 + (P - 0.3)^2 \leq 2.2$, est considérée, la zone sous contrôle est délimitée par le cercle (en vert). Cela signifie que les points à l'extérieur du cercle, mais à l'intérieur du carré formé par les limites supérieures et inférieures

1. http ://www.c2mi.ca/

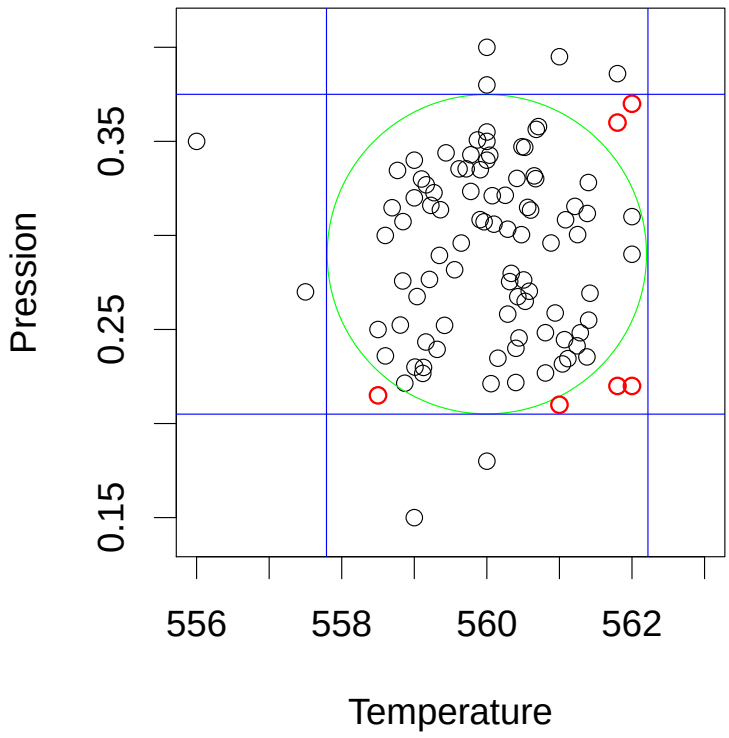


Figure 1.1 Figure illustrant les limites de contrôle d'une température et d'une pression d'un cycle Carnot sans et avec prise en compte des relations entre ces paramètres.

de chaque variable, seraient considérés à tort comme fonctionnels, si la décision est basée uniquement sur les limites monodimensionnelles. Si les limites supérieures et inférieures sont plus serrées pour éviter la non-détection de défauts, le nombre de fausses alarmes augmente. Ainsi, la prise en considération des relations entre les variables produits et/ou les paramètres contrôlés est indispensable pour une maîtrise efficace des processus. Plusieurs chercheurs se sont penchés sur la question et ont fini par proposer des méthodes de contrôle de processus multidimensionnelles. Certains auteurs ont proposé des versions multidimensionnelles des cartes de contrôle connues sont proposées, par exemple, des versions multidimensionnelles de la carte d'Hotelling, EWMA ou encore CUSUM. La majorité de ces cartes est développée sous une hypothèse assez forte qui est la normalité des données. En fait, les chercheurs supposent que le théorème centrale-limite vient valider cette hypothèse. Or, ceci n'est pas démontrable lorsque la taille des sous-groupes considérés est faible ou encore lorsque nous considérons des cartes individuelles. Pour faire face à ce problème, certains chercheurs ont proposé des cartes qui ne supposent pas la normalité des données. Néanmoins, le développement de ce type de cartes nécessite la connaissance de la distribution des données. Trouver la distribution des données n'est pas toujours évident. Les experts du domaine n'ont pas

toujours des connaissances approfondies en statistiques. Ce qui peut compliquer l'utilisation de ces cartes. Un autre type d'outils de maîtrise de processus est développé sans hypothèses ou connaissances de la distribution des données. Ce type d'outils inclut des cartes de contrôle multidimensionnelles et des outils basés sur les algorithmes d'apprentissage statistique. Ces outils tiennent en considération les relations entre les paramètres contrôlés et elles sont non paramétriques. Peu parmi les recherches axées sur les outils statistiques de maîtrise de processus se sont intéressées à fournir un support visuel à leurs outils. Ainsi, si une dérive ou un défaut est détecté, le diagnostic doit passer par un expert soutenu par des cartes de contrôle mono-dimensionnelles. Les personnes qui ne maîtrisent pas les algorithmes statistiques vont naturellement se tourner vers les supports visuels, c.a.d les graphiques. Cette remarque a été confirmée à travers les échanges que nous avons effectués avec Teledyne Dalsa. Les ingénieurs étaient plus réceptifs à des outils visuels plutôt qu'à des algorithmes statistiques. L'absence d'un support visuel facile à interpréter et à utiliser, a en quelques sortes, freiné l'utilisation des outils de contrôle multidimensionnelles. Les outils graphiques, généralement utilisés, restent quand même restreints aux outils monodimensionnels, malgré la disponibilité des données multidimensionnelles (collectées par les différents facteurs installés). Ceci peut être du a une méconnaissance des graphes multidimensionnelles ou, encore, a la difficulté de l'utilisation ou de l'interprétation de certains entre eux.

Tout outil de contrôle de processus multidimensionnel doit permettre de vérifier si le processus est sous contrôle ou pas, doit prendre en compte les relations entre les variables et paramètres et doit offrir un support pour permettre le diagnostic de tout défaut détecté(Jackson cité dans Bersimis *et al.* (2007)). La troisième propriété qui mentionne qu'un outil de contrôle doit offrir un support de diagnostic, n'est pas vérifiée par les outils actuellement disponibles. Donc, la représentation graphique des données en vue de les explorer est primordiale pour un contrôle efficace et complet. Ce travail répond, également, à un besoin identifié chez la compagnie partenaire Teledyne Dalsa. En amorçant le développement, toujours en partenariat avec Teledyne Dalsa, une seconde problématique s'est révélée, i.e. le manque de données historiques. En effet, chez Teledyne Dalsa comme chez bien d'autres compagnies, la collecte de données peut être coûteuse en termes de ressources humaines et en termes de temps. La collecte de 60 observations fonctionnelles a pris entre 4 et 5 mois et elle a nécessité la disponibilité d'un technicien chargé uniquement de la collecte pendant une certaine période. Jusqu'à la fin de cette thèse, nous n'avons pas pu obtenir plus qu'une seule observation non fonctionnelle. Ainsi, il fallait proposer un outil de maîtrise de processus qui se base sur un nombre de données historiques très limité. La collecte de données peut être difficile lors du démarrage d'une usine ou de l'acquisition d'une nouvelle machine. Le contrôle doit être réalisé en se basant sur un nombre limité de données historiques.

La question soulevée par cette thèse est comment proposer une approche efficiente de contrôle des processus et des machines. Cette question se décompose, elle-même, en 4 sous questions :

- Quel graphe multidimensionnel peut-on utiliser pour explorer les données de production dans l'objectif d'un meilleur contrôle des processus ?
- Serait-il possible d'utiliser ce type de graphe pour proposer une méthode visuelle de contrôle qui tient en compte les relations entre les différents paramètres et variables ?
- Quelles sont les caractéristiques d'un support visuel optimisé ? Comment peut en proposer un outil de contrôle fiable ?
- Serait-il possible d'utiliser ce même graphe pour proposer une méthode de contrôle dans le cas ou le nombre de données historiques est limité ?

L'objectif de cette thèse est d'exploiter la représentation graphique des données industrielles multidimensionnelles pour soutenir la de maîtrise de processus et le contrôle qualité. Il s'agit de proposer une méthode de visualisation de données multidimensionnelles et de l'optimiser pour proposer un support visuel de contrôle qualité qui vérifie les propriétés d'un outil de contrôle, c'est à dire, un outil qui ne dépend pas de la distribution des données, il tient compte des relations entre les variables et offre un support visuel de diagnostic des défauts. Nous souhaitons proposer une approche qui soutient les ingénieurs de qualité ou de procédés dans la maîtrise des processus, quand les données historiques sont disponibles en grand nombre ou en quantité limitée.

Pour atteindre cet objectif, nous proposons d'intégrer les coordonnées parallèles qui constituent un type de graphes multidimensionnelles avec les concepts de cartes de contrôle et des outils de maîtrise de processus pour proposer un outil de maitrise de processus. Le choix des coordonnées parallèles est basé sur la facilité de son utilisation par rapport aux autres types de graphes multidimensionnelles telles que les graphes Kiviat dans le cas de présence de données massives. L'outil doit offrir 2 types de cartes en fonction de la disponibilité des données historiques. La première carte est proposée dans le cas ou un nombre assez élevé de données historiques est disponible. Elle est basée sur la caractérisation de la zone de bon fonctionnement notée BOZ pour best operating zone. La deuxième carte est basée sur les graphes de densité en coordonnées parallèles lorsque peu de données historiques sont disponibles. Dans les 2 cas, le développement de ces cartes est axé sur 2 étapes principales :

- Une représentation optimisée des données en coordonnées parallèles ;
- La caractérisation de BOZ en utilisant des données historiques.

La première étape est liée aux graphes en coordonnées parallèles. En effet, les coordonnées parallèles représentent une technique assez performante pour la représentation des données multidimensionnelles. Théoriquement, il n'y a pas de limites sur le nombre d'observations ou

de variables à visualiser. Lorsque les observations sont assez nombreuses que des milliers de lignes se superposent, l'extraction de l'information fiable de ces graphiques devient impossible. En effet, Peng *et al.* (2004) mentionne qu'en coordonnées parallèles, la visualisation des relations entre les attributs n'est possible qu'entre les attributs adjacents. Nous proposons, dans cette thèse, un cadre général d'arrangement de variables qui s'adapte à l'objectif. Un exemple d'objectif est la mise en évidence des dépendances entre les variables, ou encore la séparation des données. Ce cadre se base sur la définition de 3 fonctions. Les 2 premières fonctions sont des fonctions de probabilités et elles définissent le concept de réarrangement. Par exemple, si le concept de réarrangement est la dépendance des variables, ces fonctions sont définies comme la probabilité jointe et le produit de probabilités marginales. La troisième fonction définit la statistique d'arrangement, par exemple, $u \log(u)$ pour la statistique d'information mutuelle. Le cadre d'arrangement de variables est évalué à l'aide de différentes bases de données. Chacune des bases est choisie de manière à souligner les avantages et limitations de l'approche proposée. La première base est une base de qualité de vins qui est largement utilisée, dans la littérature, pour évaluer les algorithmes d'arrangement de variables. Dans cette thèse, celle-ci est utilisée pour illustrer l'impact du changement de la statistique sur l'ordre proposé, lorsque l'objectif d'arrangement est de souligner les dépendances entre les attributs. Le deuxième test est réalisé à l'aide d'une base de données génétiques à haute dimension. Les données génétiques sont ordonnées pour souligner le concept de séparation de données et pour expliquer le comportement de l'approche lorsque les données sont de très haute dimension.

La deuxième étape est différente selon le type de cartes. Pour le premier type de cartes, la zone sous contrôle est caractérisée en déterminant une enveloppe, ou encore une limite supérieure et inférieure, de la zone de meilleur fonctionnement (BOZ pour le terme anglais, Best operating zone). L'approche utilisée est basée sur la standardisation et la visualisation en coordonnées parallèles des données historiques fonctionnelles. Ces données représentées en coordonnées parallèles sont, ensuite, projetées dans un plan Cartésien à 2 dimensions. Les coordonnées de la courbe enveloppe de la BOZ sont déterminées par rapport aux minimums et maximums des données projetées. Un pas est ajouté ou soustrait des maximums et minimums pour éviter le surajustement des cartes aux données historiques. La valeur de ce pas est définie à l'aide de la technique de validation croisée. La classification des nouvelles observations se fait, dans un premier temps, avec la BOZ. Une observation, dans cette zone, a une probabilité plus élevée d'être fonctionnelle que d'être un défaut. Une observation en dehors de cette zone est classée comme défaut. Ainsi, en se basant sur la BOZ uniquement, une observation, en dehors est un défaut, une observation à l'intérieur de la BOZ a une forte probabilité d'être fonctionnelle, mais la probabilité qu'elle soit un défaut n'est pas nulle.

Pour mieux distinguer les deux classes d'observations (fonctionnelles et défauts), la BOZ est, dans un deuxième temps, raffinée en la répartissant en plusieurs segments représentant des sous-zones de bon fonctionnement. La méthode de segmentation $k$-moyennes est implémentée. Le nombre de segments le plus adéquat est déterminé avec le critère silhouette qui vise à minimiser les variations dans un segment et à maximiser les variations entre les différents segments. Ces zones viennent réduire la probabilité de non-détection de défauts qui sont à l'intérieur de l'enveloppe de la BOZ. Ce type de cartes requiert la disponibilité d'un nombre assez élevé d'observations d'historiques fonctionnelles (300 à 1000 observations pour les cas étudiés). Une nouvelle observation qui n'appartient pas à la BOZ ou qui n'appartient pas à un segment de fonctionnement ne suit pas la structure (ou la tendance) suivie par les données historiques. La probabilité qu'une observation soit à tort classée comme défaut est la probabilité que cette nouvelle observation soit fonctionnelle bien qu'elle a un comportement différent de toutes les observations considérées lors du développement des cartes. Ces cartes sont également évaluées à l'aide de 2 bases de données différentes. La première base (base SPAM) caractérise le comportement d'un processus de formage-remplissage-scellage. L'objectif de ce test est de présenter la carte, expliquer son fonctionnement et faire un premier test d'évaluation. La deuxième base de données est une base de courriers électroniques. Il s'agit d'identifier les courriers électroniques utiles des spams qui représentent les défauts. Les cartes développées sont comparées aux cartes d'Hotelling. L'impact de l'ordre des variables sur la performance de la carte est, également, évalué. De plus, pour ce type de cartes, la distance opérationnelle moyenne (ARL) est évaluée, pour une base de données simple, à l'aide de la méthode de simulation Monte-Carlo. L'ARL est comparée à celle des cartes d'Hotelling. La facilité d'utilisation des cartes est testée à travers une expérience d'utilisations (20 utilisateurs). Dans la suite du document, nous appelons ce type de cartes, les cartes BOZ.

Pour le deuxième type de cartes de contrôle, la zone sous contrôle est caractérisée par une fonction de densité. Également, pour ce type de cartes, les données fonctionnelles représentées en coordonnées parallèles sont projetées dans un plan à 2 dimensions. Ensuite, la fonction de densité des points bidimensionnelles issus de la projection est estimée par la méthode d'estimation de densité par noyaux. Lorsque cette fonction est représentée, elle permet de distinguer les zones les plus denses des zones moins denses. Les zones les plus denses représentent les zones ou les points ont une plus grande probabilité d'être fonctionnels contrairement aux zones moins denses qui contiennent les zones hors contrôle. Ce type de cartes est développé avec un nombre très limité de données historiques. Pour garantir la fiabilité de la carte, l'échantillon de données historiques doit être classé (données fonctionnelles et défauts) avec certitude et doit être représentatif, au moins de la zone sous contrôle. Ceci est par exemple le cas pour les données de Teledyne Dalsa. Les cartes basées sur la densité

sont évaluées avec la base de données SPAM. Leur courbe d'apprentissage est tracée et est comparée à celle des réseaux de neurones et à celle des cartes d'Hotelling. L'objectif de ce test est d'évaluer la performance des cartes en fonction de la taille de la base de données historiques (base d'apprentissage).

Cette thèse commence, après l'introduction, par une revue de littérature (Chapitre 2) des outils de maîtrise des processus et de différents concepts reliés à la méthodologie de développement des cartes. La présentation des cartes de contrôle avec support graphique visuel mène à présenter quelques types de graphes multidimensionnels et en particulier, des coordonnées parallèles, graphes utilisés dans cette thèse. Il s'avère alors logique d'expliquer la dualité entre le plan bidimensionnel Cartésien et les coordonnées parallèles et de décrire, par la suite quelques études réalisées en vue d'améliorer la visualisation des données en coordonnées parallèles, i.e. les techniques d'arrangement des variables. Finalement, un état de l'art des méthodes de segmentation appliquées particulièrement aux coordonnées parallèles est présenté. Ce chapitre est terminé par un survol de quelques méthodes d'estimation des fonctions de densité et des graphes de densité en coordonnées parallèles. Le chapitre 3 présente le cadre général d'arrangement des attributs ou paramètres en coordonnées parallèles. Deux applications de ce cadre sont également présentées dans le même chapitre, soient la dépendance des attributs et la séparation des données. L'approche proposée dans ce chapitre est ensuite utilisée comme première étape dans le développement des cartes de contrôle décrit dans le chapitre 4. Ce chapitre 4 présente la méthodologie de conception des cartes BOZ et des cartes densité. Dans le chapitre 5, le cadre d'arrangement des variables et les cartes de contrôle sont évalués et discutés. Cette thèse est close par une conclusion avec un retour sur les points les plus importants des solutions proposées et par des perspectives d'amélioration. Dans la suite de la thèse, pour ne pas confondre paramètres des modèles (cartes de contrôle) avec paramètres de machines, nous utilisons le terme variables ou attributs pour paramètres de machines et paramètres pour paramètres des cartes.

# CHAPITRE 2 REVUE DE LITTÉRATURE

Dans le chapitre 1, nous avons identifié certains problèmes et limitations dans les outils de contrôle trouvés dans la littérature. Nous avons également discuté de la difficulté de l'exploitation des données de production dans l'objectif de maitrise des processus de production et du diagnostic des défauts et dérives identifiés. Nous avons introduit une méthodologie qui utilise les bases des outils de contrôle et de cartes de contrôle combinées avec un outil de visualisation des données multidimensionnelles qui est les graphes en coordonnées parallèles. Dans ce chapitre, une revue de littérature en lien avec la problématique de recherche et avec la méthodologie suivie est présentée. La revue de littérature présentée regroupe différents axes qui ne sont pas d'habitude présentés dans un même document. Même s'ils ne sont pas directement reliés, leur compréhension est nécessaire pour comprendre la méthodologie du développement de l'outil de maîtrise de processus proposé dans cette thèse. La section 2.1 présente une revue des outils de contrôle de processus, de détection de défauts et particulièrement des cartes de contrôle. multidimensionnelles. Les supports visuels aux outils de contrôle sont également, présentés dans la section 2.1.2. Quelques critères de performance utilisés pour évaluer les cartes de contrôle sont discutés dans la section 2.1. La section 2.2 décrit certaines techniques de visualisation des données multidimensionnelles avec un accent sur les coordonnées parallèles, graphes utilisés pour concevoir l'outil de contrôle proposé dans ce document. Dans la même section, différents aspects liés aux coordonnées parallèles sont présentés. La sous-section 2.2.2 explique la dualité entre les graphes Cartésiens orthogonaux et les coordonnées parallèles. La section 2.2.3 présente une revue de littérature des techniques d'arrangement d'attributs, en coordonnées parallèles. La sous-section 2.2.4 parcourt quelques techniques de segmentation et leur application aux coordonnées parallèles. Le chapitre 2 se termine par un état de l'art des graphes de densité appliqués, particulièrement, aux coordonnées parallèles.

## 2.1 Outils de contrôle multidimensionnels

### 2.1.1 Revue des outils de maîtrise des processus

Les processus de production sont composés de nombreux sous-systèmes, très interconnectés et interactifs. Ceci est le cas des usines de semi-conducteurs. Pour garantir la qualité du produit final, la maîtrise des processus de fabrication est indispensable. La prise en compte des relations entre les variables du produit et les variables de l'équipement est nécessaire pour

un contrôle efficace. Cela signifie que le contrôle séparé de chaque variable peut conduire à un contrôle incomplet ou inefficace tel qu'expliqué dans le chapitre 1. Par conséquent, les techniques de contrôle de processus multidimensionnels sont proposées. En 1947, Hotelling (cité dans (Lowry et Montgomery, 1995)) introduit une technique de contrôle de processus multidimensionnels. Il s'agit d'une version multidimensionnelle des cartes de contrôle de type Shewhart. Les cartes de Shewhart ne sont pas efficaces pour détecter de petits changements dans les variables du processus. Ensuite, des versions multivariées d'autres cartes de contrôle connues ont été proposées. Crosier (1988) suggère une généralisation des cartes de contrôle CUSUM. Lowry et Rigdon (1992) présente une version multivariée de la carte EWMA. Ces deux dernières cartes sont plus efficaces avec les petites variations que les cartes de Shewhart. Elles considèrent la matrice de variance-covariance tout en définissant les limites de contrôle autour de la cible (définie assez souvent comme la moyenne). Malgré le progrès des recherches dans la proposition d'outils de contrôle de processus multivariés, la carte $T^2$ d'Hotelling reste l'une des méthodes multivariées les plus implémentées dans l'industrie. La carte d'Hotelling est, ainsi, utilisée comme carte de référence pour évaluer et comparer les solutions proposées dans cette thèse. Dans la revue de littérature, nous mettons, en particulier, l'accent sur l'explication de cette carte. Pour plus de détails sur d'autres cartes de contrôle multidimensionnelles, Lowry et Montgomery (1995), Mason et Young (1997) et Montgomery et LAWRENCE (1995) constituent de bons supports. Pour les cartes d'Hotelling, comme pour plusieurs autres cartes de contrôle, nous pouvons distinguer des cartes qui vérifient une statistique d'un sous-groupe d'observations, comme la moyenne et des cartes pour des observations individuelles. Nous nous intéressons aux cartes individuelles.

La conception des cartes d'Hotelling se fait en 2 phases. Une première phase (phase I) qui consiste à sélectionner les données historiques de confiance qui vont servir à établir les limites et classer les nouvelles observations dans la phase II :

— Phase I : une statistique est définie pour chaque vecteur représentant une observation historique $\mathbf{x}$.

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}}_h)\mathbf{\Sigma}_h^{-1}(\mathbf{x} - \bar{\mathbf{x}}_h)$$

ou $\bar{\mathbf{x}}_h$ et $\mathbf{\Sigma}_h$ sont, respectivement, le vecteur de moyenne échantillonnale et la matrice de variance-covariance estimés à partir des $n$ données historiques. Les limites de contrôle sont définies comme suit :

$$UCL = \frac{(n-1)^2}{n}\beta_{1-\frac{\alpha}{2},\frac{p}{2},\frac{n-p-1}{2}}$$

$$LCL = \frac{(n-1)^2}{n}\beta_{\frac{\alpha}{2},\frac{p}{2},\frac{n-p-1}{2}}$$

$\beta$ est la quantile de la distribution Bêta et $\alpha$ est le niveau d'incertitude. Nous continuons la

phase II avec les données historiques qui vérifient la condition $LCL \leq T^2 \leq UCL$.

— Phase II : cette phase permet de classer les nouvelles observations.

Nous définissons la statistique $T^2$ pour les nouvelles observations, représentées par les vecteurs $\mathbf{o}$ comme suit :

$$T^2 = (\mathbf{o} - \bar{\mathbf{x}})\mathbf{\Sigma}^{-1}(\mathbf{o} - \bar{\mathbf{x}})$$

où $\bar{\mathbf{x}}$ et $\mathbf{\Sigma}$ sont les estimateurs de la moyenne des données et la matrice de variance-covariance estimés à l'aide des données historiques retenues à la fin de la phase I.

$$\bar{\mathbf{x}} = (\overline{x_1}, \overline{x_2}, .., \overline{x_p}) \text{ où } \overline{x_i} = \frac{\sum_{j=1}^{n} x_{ij}}{n} \, ; \forall i \in \{1, .., p\}$$

$$\mathbf{\Sigma} = (\sigma_{ij})_{1 \leq i \leq p, 1 \leq j \leq n} \text{ où } \sigma_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j}) \, ; \forall i \in \{1, .., n\} \text{ et } \forall j \in \{1, .., p\}$$

Lorsque l'hypothèse nulle $H_0 : \mu = \bar{\mathbf{x}}$ et $\mathbf{S} = \mathbf{\Sigma}$ est valide, ou $\mu$ et $\mathbf{S}$ sont la moyenne et la matrice de variance-covariance de la population,

$$\frac{n(n-p)}{p(n+1)(n-1)} T^2 \sim F_{p,n-p},$$

où $F_{p,n-p}$ est une distribution de Fisher avec $p$, $n-p$ degrés de liberté. Le niveau d'incertitude toujours fixé à $\alpha$, les limites inférieure et supérieure sont définies comme suit :

$$UCL = \frac{(n+1)(n-1)p}{n(n-p)} F_{1-\frac{\alpha}{2},p,n-p}.$$

$$LCL = \frac{(n+1)(n-1)p}{n(n-p)} F_{\frac{\alpha}{2},p,n-p}.$$

Chaque fois qu'une nouvelle observation doit être surveillée, la distance $T^2$ doit être calculée et comparée à la limite supérieure $UCL$ et à la limite inférieure $LCL$.

La carte de contrôle d'Hotelling multidimensionnelle est l'une des cartes de contrôle multidimensionnelles les plus utilisées. Cependant, comme beaucoup d'autres cartes de contrôle, incluant les cartes multivariées EWMA et multivariées CUSUM, elle est sous une hypothèse très restrictive, c'est que tous les processus considérés suivent une distribution normale multivariée. Cela montre des limitations dans les cas d'application de problèmes réels de contrôle de qualité (Gang *et al.*, 2013). En fait, les praticiens ont parfois l'impression que le théorème de central limite «viendra régler» et assurera d'une manière ou d'une autre l'efficacité attendue des cartes. Bien que cela soit vrai pour certaines cartes de contrôle basées sur des moyennes de certaines statistiques à partir de processus sous contrôle, il est loin d'être vrai en général, spécialement pour les cartes de contrôle individuelles ou qui considèrent des petits sous-groupes. Certains chercheurs ont essayé de traiter le problème causé par la non-normalité

des données en développant certains outils de contrôle sans hypothèse sur la distribution des données, également appelés outils de contrôle non paramétriques. Par exemple, Sun et Tsung (2003) introduisent une carte de contrôle multivariée basée sur la distance par rapport au noyau, qui utilise des méthodes à support vectoriel motivées par la théorie de l'apprentissage statistique. Ses limites de contrôle ne dépendent pas de la distribution, mais s'adaptent aux données réelles. Cette méthode est plus flexible car elle n'est pas valide uniquement dans le cas de données normales. Chen *et al.* (2011) proposent également d'utiliser les machines à supports vectoriels pour construire un modèle de détection de défauts dans une usine d'énergie thermique. Leur modèle montre un taux de classification correcte de 93.13% comparé à 86.5% pour les réseaux bayésiens et 76.71% pour l'analyse en discriminants linéaires. Qiu et Hawkins (2003) transforment les cartes CUSUM multivariées pour en générer une version non paramétrique. Cependant, cette carte suppose que la distribution des données est, quand même, connue. Dans la pratique, les distributions réelles des données sont généralement méconnues. D'un autre coté, l'estimation des fonctions de distributions peut s'avérer complexe pour les industriels qui ne sont pas forcément familiers avec les outils statistiques comme mentionné dans le chapitre 1. C'est la raison principale derrière le développement des méthodes non paramétriques. Bakir (2004) propose une carte de type Shewhart qui est basée sur l'association de classes aux observations groupées. Les avantages des cartes de contrôle non paramétriques incluent leur applicabilité à des données nettement non normales, leur robustesse par rapport aux valeurs aberrantes, leur longueur moyenne de séquence constante (appelée souvent ARL pour average run length) et les taux de fausses alarmes en contrôle. En particulier, cette robustesse de distribution pourrait constituer un avantage significatif dans les situations de démarrage de production où nous n'avons généralement pas connaissance de la distribution sous-jacente. Qiu (2008) propose de transformer, dans un premier lieu, chaque mesure d'équipement en une variable binaire, qui est une fonction indicatrice de l'événement ; la mesure est plus grande que sa médiane lorsque le processus est sous contrôle. Ensuite, un modèle log-linéaire est utilisé pour décrire les associations possibles entre les variables binaires, fournissant un estimateur log-linéaire de la distribution conjointe des mesures lorsque le processus est sous contrôle. Ceci permet de caractériser la zone de fonctionnement.

Gang *et al.* (2013) développent une méthode basée sur la région de la plus probable pour assurer que tous les points sous contrôle ont une plus grande possibilité d'occurrence que les points hors contrôle. La région la plus probable désigne la zone de confiance. La méthode proposée satisfait aux exigences relatives aux fausses alarmes et elle est précise même avec des processus de distribution non normale. Cette méthode est performante pour les distributions asymétriques ou multimodales, où les limites de contrôle traditionnelles de $T^2$ et les limites de contrôle de $T^2$ estimées par les méthodes de noyau sont toutes les deux très inexactes

(Umit et Cigdem, 2001). Plusieurs auteurs ont étudié l'utilisation de modèles d'intelligence artificielle et particulièrement les réseaux de neurones artificiels ((Jakubek et Strasser, 2004) et (Gonzaga *et al.*, 2009)). Yu *et al.* (2008) emploient l'algorithme génétique pour extraire des règles d'association qui expriment les relations de causalité entre les attributs des processus et les mesures des produits ou l'état du processus de production. Ce modèle a démontré une haute précision dans la maîtrise des signaux anormaux des systèmes de production.

Certains auteurs suggèrent l'utilisation de techniques de sélection de variables combinées avec les méthodes de contrôle de processus multivarié. Par exemple, Zou et Qiu (2012) adaptent la méthode de sélection des variables LASSO au problème de contrôle statistique des processus pour proposer une nouvelle méthode de contrôle. Bleakie et Djurdjanovic (2013) développent une technique pour la maîtrise des processus dans les usines de semi-conducteurs. Ils implémentent une analyse à discriminant linéaire pour sélectionner les attributs dynamiques qui sont les plus sensibles à la modification des recettes machines. Ensuite, les distributions des attributs dynamiques sont estimées à l'aide de mélanges de Gaussiennes exprimant, ainsi, les changements de ces attributs lorsque les recettes changent.

Un autre type de contrôle de processus statistique multidimensionnel est basé sur la projection de variable latente, telle que l'analyse en composantes principales (PCA) et la méthode des moindres carrés partiels (PLS). L'objectif de l'analyse des composantes principales est de réduire la complexité des données. Si un grand nombre de facteurs (composantes) est nécessaire pour représenter les variables, il devient très peu nécessaire d'effectuer une analyse en composantes principales. L'analyse des composantes principales a pour but de réduire le nombre de variables ou attributs utilisés pour développer le modèle de maîtrise de processus. Jiang *et al.* (2016) intègrent l'analyse en composantes principales avec les réseaux bayésiens pour proposer un modèle de détection de dérives. Les modèles basés sur les composantes principales n'établissent pas un lien direct entre les variables de sortie et les attributs d'entrée. Il peut être, alors, difficile d'attribuer une interprétation significative aux composantes principales. Le diagnostic des défauts ou dérives détectés n'est, alors, pas intuitif. Par conséquent, les cartes de contrôle des composantes principales peuvent compléter, mais non remplacer, les cartes multidimensionnelles. Nomikos et MacGregor (1994) appliquent l'analyse en composantes principales pour établir une carte de contrôle multidimensionnelle. Kourti et MacGregor (1995) proposent les méthodes de projection multi-blocs multi-manières pour extraire des informations pertinentes sur les données de configuration de lots et les données de trajectoires multivariées. La projection se fait sur des espaces de dimensions réduites définis par les variables latentes ou les composantes principales. Chen et Liu (2002) utilisent l'analyse en composantes principales et la méthode des moindres carrés partiels dynamiques pour contrôler les variables de processus et les variables de qualité. Weese et Jones-Farmer

(2016) citent certaines techniques de classification telles que les $k$ plus proches voisins et les techniques à supports vectoriels. Pour une vue d'ensemble complète sur les cartes de contrôle non paramétriques, nous pouvons nous référer à Bakir (2001), Chakraborti *et al.* (2001) et Weese et Jones-Farmer (2016). Bect *et al.* (2015) combinent plusieurs techniques de data mining et d'apprentissage statistique ; l'analyse en composantes principales, la segmentation par la méthode des $k$ moyennes, l'analyse en discriminants canoniques pour caractériser le fonctionnement sous contrôle d'un système en utilisant des données historiques. Le développement de leur modèle est réparti en 2 étapes. La première consiste en la caractérisation de la zone de fonctionnement sous contrôle et la deuxième consiste en l'identification de la position des points non fonctionnels par rapport au comportement sous contrôle. Venkatasubramanian *et al.* (2003) revoient des techniques de maîtrise de processus ou détection de défaut et constituent une bonne source d'information. Precup *et al.* (2015) étudient certaines techniques d'apprentissage statistique et de data mining telles que les machines à supports vectoriels et les classificateurs à logique floue. En revenant aux caractéristiques des outils de contrôle citées dans la section 1, plusieurs outils valident les 2 premières caractéristiques, soient elles permettent de vérifier si le processus sous contrôle ou pas et ils prennent en compte les relations entre les attributs contrôles. Majoritairement, les relations considérées sont les corrélations. La troisième condition qui consiste à fournir un diagnostic aux problèmes détectés, est communément omise comme expliqué précédemment. La sous-section 2.1.2 présente un bref aperçu de ce champ, c'est-à-dire des implémentations visuelles des tableaux de contrôle.

### 2.1.2 Cartes de contrôle multidimensionnelles à support visuel

Les cartes de contrôle monodimensionnelles sont assez souvent accompagnées d'un support visuel. Elles sont représentées sous forme de graphiques monodimensionnels montrant la mesure cible qui est représentée, dans plusieurs cas, par la moyenne ou la médiane de la statistique utilisée. Les cartes monodimensionnelles montrent également, la limite supérieure et inférieure de la statistique visualisée pour délimiter la zone de contrôle comme illustré par la figure 1.1. Tous les points en dehors de ces limites sont considérés comme des défauts ou des dérives. Les limites de contrôle sont dans plusieurs cas définies par une distance autour de la cible. Les limites sont fréquemment définies à $\pm\sigma$ ou $\pm 3\sigma$ ou $\sigma$ est l'écart type du paramètre monitoré (Bersimis *et al.*, 2007). Ces limites sont développées pour identifier la zone fonctionnelle, mais elles ne sont pas précises, comme elles ne prennent pas en compte les relations entre les variables comme expliquées dans le chapitre 1.

L'idée des cartes de contrôle multidimensionnelles est similaire. Une carte de contrôle multi-

dimensionnelle doit visualiser la zone de points fonctionnels avec ses limites. La carte multidimensionnelle doit permettre le diagnostic des défauts ou dérives en vérifiant un seul graphique. Cependant, quoique les outils de contrôle multidimensionnels ont fait l'objet de plusieurs recherches, la visualisation demeure un axe à explorer. L'implémentation de supports visuels nécessite l'utilisation de graphes multidimensionnels. Quelques études ont proposé dans l'objectif de fournir un support visuel aux outils de contrôle. La visualisation des limites multidimensionnelles qui tiennent en compte les dépendances entre les variables n'est pas triviale. Umit et Cigdem (2001), dans leur revue de littérature mentionnent 2 types de graphiques de visualisation des cartes de contrôle, soient, les starplots et les cartes à profil multivariées. Les coordonnées parallèles sont, également, utilisées par certains auteurs. Une explication des graphes en coordonnées parallèles est présentée dans la section 2.2. Brooks *et al.* (2004) proposent une nouvelle méthode pour gérer les alarmes en utilisant les coordonnées parallèles. Ils suggèrent de délimiter la BOZ à partir des données historiques. Ils utilisent des méthodes géométriques pour déterminer la courbe enveloppe de la BOZ. Ensuite, lorsqu'un paramètre est fixé, de nouvelles limites sont recalculées pour tous les autres attributs pour que le point en question reste sous contrôle. Leur méthode démontre des résultats assez intéressants avec un taux de fausses alarmes égal à 10% pour le cas du processus chimique de IneosChlor and Mallinckrodt. La méthodologie de développement de leurs cartes de contrôle reste quand même assez floue et très peu expliquée. Ceci est probablement dû au fait qu'ils ont utilisé le même algorithme comme solution logiciel. Albazzaz *et al.* (2005) utilisent les coordonnées parallèles comme outil de contrôle qualité. Ils représentent les données historiques en coordonnées parallèles. Les données qui n'appartiennent pas à la zone dense sont considérées comme clairement non fonctionnelles. Ces valeurs aberrantes sont retirées de la base de données. Les données restantes sont, alors, représentées, une deuxième fois, pour enlever les données probablement non fonctionnelles qui sont à leur tour enlevées. Les données sont visualisées une dernière fois pour identifier les données probablement fonctionnelles. Ainsi, les données restantes représentent la BOZ. Albazzaz et Wang (2006) suggèrent de transformer les données non normales en données suivant une loi normale (Gaussienne). Ceci permet de rapprocher les valeurs aberrantes. Ensuite, ils proposent appliquer l'analyse en composantes principales pour réduire le nombre d'attributs. Ensuite, les limites de contrôle ont été déterminées séparément pour chaque composante. Les limites sont fixées à $\pm\sigma$ ou $\pm 3\sigma$ autour de la moyenne. Visualisée en coordonnées parallèles, cette technique ne conserve pas toutes les informations puisque l'objectif des composantes indépendantes est de réduire le nombre de variables. Dunia *et al.* (2013) utilisent, aussi, les coordonnées parallèles. Ils proposent 3 types de zones opérationnelles nommées région de confiance. Ces zones sont soient rectangulaires, elliptiques et rectangulaires pivotées dans un espace Cartésien à 2 dimensions. Ensuite, Du-

nia *et al.* (2013) utilisent les équations de dualité entre l'espace Cartésien et les coordonnées parallèles pour trouver les régions équivalentes en coordonnées parallèles. Cette méthode a été appliquée aux composantes principales adjacentes. Un exemple de région opérationnelle telle que définie par Dunia *et al.* (2013) est illustrée par la figure 2.1.

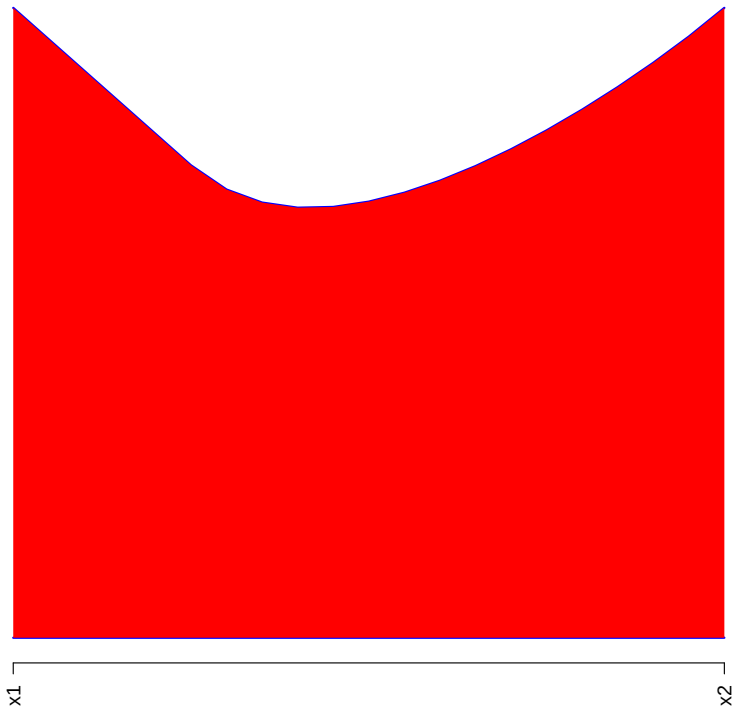


Figure 2.1 Figure illustrant une région opérationnelle elliptique représentée en rouge en coordonnées parallèles.

Une autre approche de détection de défauts (dérives) est proposée basée sur les diagrammes en étoile appelés également diagramme de Kiviat (Wang *et al.*, 2015). Cette méthode est similaire à la méthode de Albazzaz *et al.* (2005), qui suggère d'enlever d'une façon itérative, les données clairement non fonctionnelles, probablement non fonctionnelles et les données probablement fonctionnelles, mais avec une technique de visualisation différente. Finalement, Gajjar et Palazoglu (2016) intègrent les coordonnées parallèles avec un processus de contrôle qui se repose sur la visualisation des composantes principales. Comme dans Albazzaz et Wang (2006), les limites sont déterminées séparément pour chaque composante pour objectif de détection de défaut. Leur méthode, testée sur le processus "Tenesse Eastman", montre des résultats intéressants pour la détection de défauts et aussi pour le diagnostic. Le taux de classification incorrecte est réduit significativement avec leur méthode comparée à la carte $T^2$ d'Hotelling. Dans cette section, nous présentons quelques études faites dans l'objectif de

la visualisation des outils de contrôle des processus à l'aide des graphes multidimensionnels notamment des coordonnées parallèles. Il vient de soi, alors, d'expliquer de certaines techniques de visualisation de données multidimensionnelles, particulièrement, les coordonnées parallèles qui sont, par la suite, utilisées dans la conception de l'outil de contrôle proposé dans cette thèse, dans la section 2.2. Avant de présenter ceci, nous présentons quelques critères utilisés pour évaluer les performances des cartes de contrôle.

### 2.1.3 Critères de performance des cartes de contrôle

Certaines mesures de performance sont suggérées pour évaluer les cartes de contrôle proposées, les comparer et choisir celle qui s'adapte mieux au contexte. Les critères de performance les plus utilisés sont la longueur moyenne de séquence (ARL pour average run length), le temps moyen jusqu'à la première alarme et le nombre moyen d'observations jusqu'au premier signal d'alarme. L'ARL réfère au nombre d'observations visualisées avant de détecter un défaut. Nous distinguons l'ARL sous contrôle et l'ARL hors contrôle. Pour certaines cartes, l'ARL peut être calculée précisément et théoriquement. Si cela n'est pas possible, l'ARL est simulé par la méthode de Monte Carlo ou par les chaînes de Makov. Le nombre d'essais Bernouilli jusqu'à l'obtention un signal de détection de défaut, est le nombre moyen d'observations avant l'obtention d'un signal d'alarme. Il est proportionnel au temps moyen avant obtention d'un signal d'alarme (Woodall et Driscoll, 2015).

## 2.2 Graphes en coordonnées parallèles

### 2.2.1 Présentation des graphes en coordonnées parallèles

Lorsque les données sont de hautes dimensions, la représentation de chaque variable sur un graphique séparé peut aboutir à une vue d'ensemble incomplète et à une analyse inefficace. Ceci inclut le cas ou les graphiques sont utilisés pour le contrôle qualité et la maîtrise des processus. Les graphes couramment utilisés sont limités à la visualisation d'un maximum de 3 variables simultanément (Plan à 3 dimensions). La quantité de données à hautes dimensions augmente d'une façon remarquable. Les graphes multidimensionnels sont, alors, proposés, pour permettre la représentation simultanée de plusieurs variables. Ces graphes facilitent la manipulation de ce type de données. Les matrices de nuages de points, les glyphs, les graphes polaires et les coordonnées parallèles sont des exemples d'outils de visualisation de données multidimensionnelles. Dans cette thèse, une analyse approfondie des graphes en coordonnées parallèles est présentée vu qu'elles sont utilisées dans la conception de la solution proposée. En effet, les coordonnées parallèles constituent un outil puissant et facile à manipuler et à interpréter. Les coordonnées parallèles ont été, pour la première fois, proposées par D'Ocagne (1885) comme alternative aux graphes d'un espace Cartésien. Les coordonnées parallèles permettent la visualisation de données multidimensionnelles dans un espace Cartésien sans être amené à effectuer des calculs supplémentaires (Woodward et Goldsmith, 1964). Les coordonnées parallèles ont été, ensuite, davantage explorées par Inselberg et Dimsdale (1990).

Les données sont représentées par un ensemble de lignes qui passent à travers des axes parallèles représentant les variables. Ainsi, chaque point est représenté par un ensemble de lignes appelé polylignes. La figure 2.2 illustre 2 points $A(1, 1, 2.5)$ et $B(-1, 0, 1.5)$ en 3 dimensions et en coordonnées parallèles.

Les coordonnées parallèles sont intégrées dans plusieurs logiciels et outils. Dans XMDV-Tool et XDAT, l'utilisateur est capable de visualiser des données multidimensionnelles en coordonnées parallèles et de les manipuler d'une façon interactive. Ces outils permettent de représenter une matrice de données tout en accédant à une liste de fonctionnalités et d'options tels que l'application de filtre, la sélection de données, la segmentation ou encore l'arrangement des axes. Ces logiciels permettent, aussi, d'afficher le nombre total d'observations, le nombre d'observations sélectionnées, le minimum et le maximum de chaque variable. De plus, plusieurs langages de programmation statistique offrent la possibilité de visualiser des données en coordonnées parallèles. *R* permet de visualiser les données en coordonnées parallèles à partir de 2 packages, soient MASS et ggparallel. Matlab, Statistica et Python permettent, également de le faire. Plusieurs recherches sont faites en lien avec coordonnées parallèles.

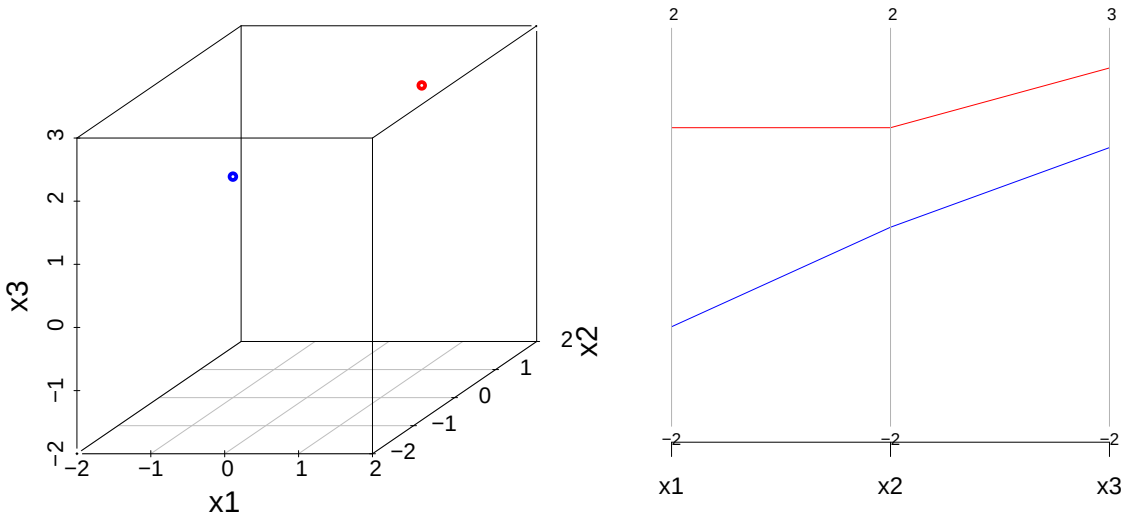


Figure 2.2 Figure illustrant la dualité entre un graphe à 3 dimensions en plan Cartésien et en coordonnées parallèles pour 2 points $A(1, 1, 2.5)$ et $B(-1, 0, 1.5)$.

Elles visent à faciliter la lecture de ces graphes et améliorer et soutenir leur exploitation. Ces recherches visent généralement à souligner les relations entre les variables, à découvrir les formes et modèles observés ou à analyser la répartition des données. Les sections suivantes en présentent quelques unes. Pour pouvoir explorer les graphes en coordonnées parallèles, il va de soi de comprendre l'interprétation des fonctions couramment utilisées en coordonnées parallèles. La dualité entre les graphes bidimensionnels du plan Cartésien et les graphes en coordonnées parallèles est étudiée dans la section 2.2.2. Le réarrangement des variables qui est un concept clef pour une représentation efficace et optimisée des coordonnées parallèles est discuté dans la section 2.2.3. Vu la démarche suivie pour le développement des cartes de contrôle proposée dans cette thèse, quelques algorithmes de segmentation sont décrits avec une application aux coordonnées parallèles dans la section 2.2.4. Cette section est close par les graphes de densité appliqués aux coordonnées parallèles.

### 2.2.2 Dualité $2D$ et coordonnées parallèles

Le choix du système de représentation, c'est-à-dire du type de graphique utilisés, détermine les formes observées pour une fonction donnée. Ainsi, il est primordial de savoir lire et comprendre les formes et modèles aperçus.

Plusieurs auteurs ont étudié la dualité entre le plan Cartésien et les coordonnées parallèles.

Ils ont expliqué la projection d'un point d'un plan orthogonal à un plan en coordonnées parallèle. Inselberg et Dimsdale (1990) et Heinrich et Weiskopf (2013) expliquent le passage d'un plan Cartésien à un plan en coordonnées parallèles. Soit un plan de 2 coordonnées parallèles $(x_1, x_2)$ comme dans la figure 2.3. Les axes des coordonnées parallèles ont pour abscisses 0 et $d$ ($d = 1$ dans cet exemple), dans le plan Cartésien, ce qui veut dire que les axes sont distants d'une distance $d$, non nulle, sinon les axes seraient confondus.

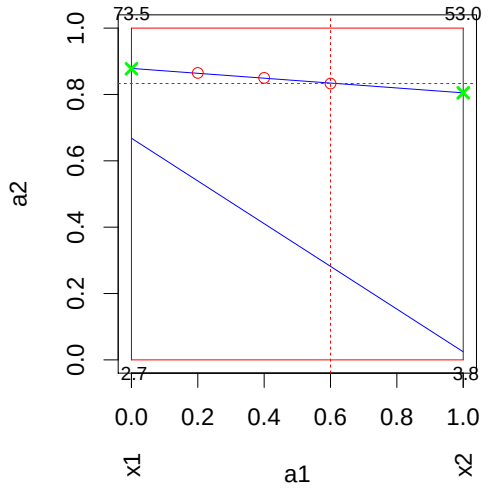


Figure 2.3 Projection d'observations en coordonnées parallèles dans un plan Cartésien.

Soit un point $A(x_1, x_2)$ dans un plan Cartésien à 2 dimensions. Ce point est représenté en coordonnées parallèles par une ligne entre le point $(0, x_1)$ et $(d, x_2)$ d'équation :

$$a_2 = \frac{x_2 - x_1}{d} a_1 + x_1$$

dans le plan Cartésien ou $a_1 \in [0, d]$. Dans la figure 2.3, ce point est représenté par la ligne bleue reliant les points en vert. Généralement, nous choisissons $d = 1$ pour des fins de simplification, comme dans l'exemple présenté. Ainsi, une fonction $f$ définie comme suit :

$$\begin{array}{rccc} f & : & E & \rightarrow & F \\ & & x_1 & \mapsto & f(x_1) \end{array} \tag{2.1}$$

tels que $E$ et $F$ sont deux sous-espaces de $\mathbb{R}$, est traduite par un ensemble de lignes d'équation :

$$a_2 = \frac{f(x_1) - x_1}{d} a_1 + x_1 \tag{2.2}$$

Pour une fonction linéaire d'équation

$$f \;:\; x_1 \;\mapsto\; \alpha x_1, \tag{2.3}$$

l'équation 2.2 qui donne les équations des droites représentant les observations et projetées dans l'espace Cartésien devient :

$$a_2 = \frac{\alpha x_1 - x_1}{d} a_1 + x_1.$$

Si, de plus, nous supposons que la distance entre les axes parallèles est $d = 1$, alors, l'équation 2.2 devient :

$$a_2 = (\alpha x_1 - x_1) a_1 + x_1.$$

Visuellement, un ensemble de points situés sur une droite dans un espace Cartésien est transformé en un ensemble de lignes qui se coupent en un point défini. La position horizontale dépend de la pente de la fonction. Si la pente est négative, le point d'intersection se trouve entre les 2 axes parallèles. Si elle est positive et inférieure à 1, le point d'intersection se trouve à la droite des 2 axes. Si elle est positive, mais supérieure à 1, ce point est à gauche des 2 axes parallèles. Cependant, étant donné que la plupart des logiciels qui permettent de représenter les coordonnées parallèles standardisent les données avant de les représenter pour éviter la compression (skewness) de certaines variables par rapport à d'autres, les points d'une droite à pente positive sont toujours transformés en un ensemble de lignes parallèles. Ceci est illustré par la figure 2.4.

La figure 2.4 montre une fonction linéaire à coefficient négative et positive, avec et sans bruit et aussi une fonction constante, dans un plan Cartésien et en coordonnées parallèles. La fonction constante est visualisée par un ensemble de lignes qui se joignent en un seul point sur le deuxième axe. Ceci est équivalent à une fonction linéaire à pente nulle. Lorsqu'un bruit blanc est ajouté à la fonction linéaire, les lignes ne sont pas parfaitement parallèles, mais la structure demeure visible. Cette structure doit disparaître lorsque le bruit augmente.
Certaines fonctions non-linéaires sont illustrées par la figure 2.5.

La figure 2.5 illustre certaines fonctions dans un plan Cartésien et en coordonnées parallèles. Les fonctions en coordonnées orthogonales peuvent être projetées en coordonnées parallèles en utilisant les équations de dualité. Les formes ou les courbes observées dans les graphiques montrent qu'une fonction quadratique sur un intervalle donné est représentée par un ensemble de lignes délimitées par une ligne et une hyperbole. Leurs positions dépendent du coefficient du terme du deuxième degré. Quand le coefficient du degré le plus élevé est positif, l'hyperbole est en haut et la ligne est en bas. Quand le coefficient est négatif, le contraire se produit.

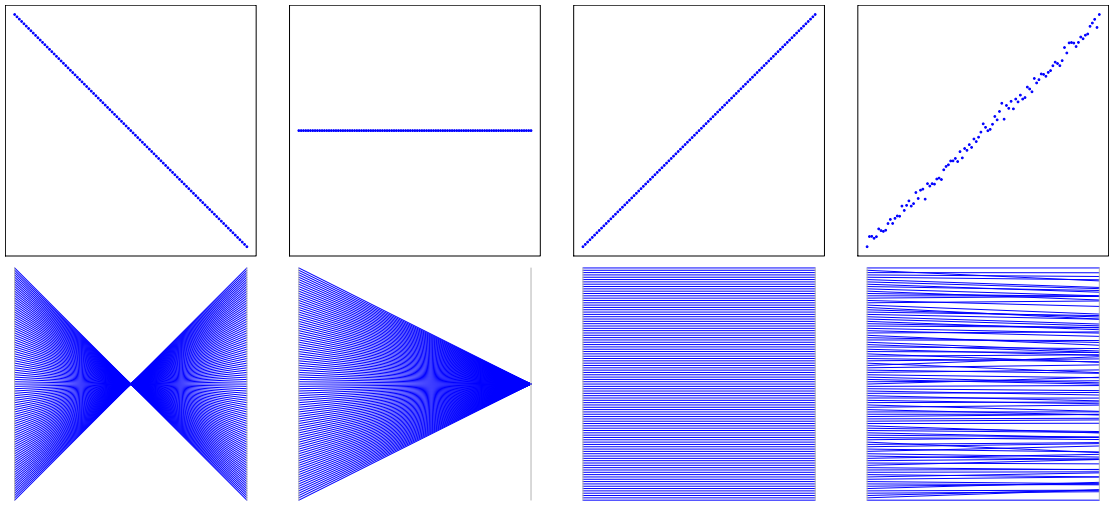

Figure 2.4 Figure illustrant la dualité entre un plan Cartésien à 2 dimensions et les coordonnées parallèles pour les fonctions linéaires. De gauche à droite, nous avons, une fonction linéaire pente négative, une fonction constante et une fonction à pente négative avec et sans bruit. Les figures d'en haut sont représentées dans un plan Cartésien. L'équivalent en coordonnées parallèles est représenté dans les figures en bas.

Généralement, les formes détectées dépendent du signe du coefficient du terme le plus élevé. Une fonction cubique peut être traduite par des lignes qui se croisent en différents points. L'ensemble d'observations est délimité par 2 lignes parallèles si le coefficient du plus grand terme est positif et par 2 courbes si le coefficient du terme au cube est négatif.

Sur la figure 2.5, une fonction sinusoïdale est également représentée. En coordonnées parallèles, la fonction sinusoïdale est semblable à la fonction cubique à coefficient négatif, mais avec des lignes qui la délimitent plutôt que des courbes. Sur la figure 2.6, nous pouvons aussi voir la fonction $f(x) = \frac{1}{x}$, cette fonction est représentée par deux courbes et une zone intérieure vide. Cette structure illustre le cas ou $x$ varie de $[-1, 1]$. Cependant, si $x$ est strictement positif ou strictement négatif, seule la partie supérieure ou inférieure de la courbe sont représentées. Ainsi, les formes qui apparaissent dépendent également du signe des attributs.

Les formes restent quand même pas si évidentes à interpréter. Plusieurs formes se ressemblent d'une fonction à une autre, par exemple, la fonction cubique et la fonction sinusoïdale sont toutes les deux représentées par des lignes qui se croisent limitées par deux courbes. La différence consiste dans le point d'intersection et dans la forme des courbes enveloppes. Malgré cette incertitude dans l'interprétation des fonctions, il est clair que lorsque 2 variables sont dépendantes, la représentation de la fonction en coordonnées parallèles montre des formes claires même lors de la présence d'un petit bruit. Ce bruit peut être encore plus présent dans le cas des données réelles compliquant ainsi l'interprétation des formes observées.

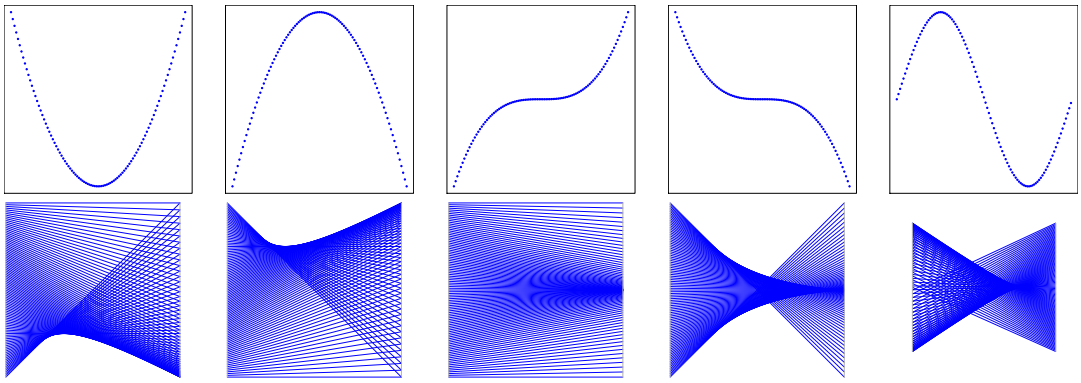

Figure 2.5 Figure illustrant la dualité entre $2D$ et coordonnées parallèles pour des fonctions non-linéaires. De gauche à droite sont représentées les fonctions suivantes : fonction carré à coefficient positif puis négatif, fonction cubique à coefficient positif et négatif et une fonction périodique (sinusoïdale).

Or, l'exploration des données peut se faire en analysant les structures observées ou en appliquant des algorithmes de segmentation. En effet, la segmentation étant une tache d'exploration de données assez intéressante et assez utilisée, la perception des segments en coordonnées parallèles est importante pour aller vers l'amélioration de leur qualité. La figure 2.7 illustre différents cas de données bidimensionnelles. Les différents cas illustrés représentent, soient des données séparables non corrélées, soient des données corrélées non séparables, soient des données corrélées et séparables ou des données non corrélées et non séparables, en coordonnées parallèles et en coordonnées orthogonales d'un plan Cartésien. Lorsque les données sont séparables non corrélées, les segments sont distinguables, mais les données ne montre pas de forme en particulier, toutefois elles sont regroupées dans des régions différentes relativement au segment. Lorsque les données sont corrélées et séparables. la forme des données de chaque segment est, de plus, claire. Les données non séparables sont représentées par des lignes toutes mélangées indépendamment du segment, mais les lignes sont presque toutes parallèles montrant la linéarité. Finalement, lorsque les données sont non corrélées et non séparables, un mélange de points apparait, les segments sont non séparables et aucune structure n'est visible.

Ainsi, cette sous-section présente une analyse de la dualité entre le plan Cartésien et les coordonnées parallèles. L'emphase est mise sur la visualisation, pour faciliter la compréhension et l'interprétation de l'outil développé. Or, tout ce qui a été détecté de relations, formes, séparabilité est visible uniquement entre 2 variables adjacentes, d'où, l'importance du choix de l'ordre de visualisation des données. L'état de l'art des algorithmes de réarrangement des coordonnées parallèles est présenté dans la sous-section 2.2.3.

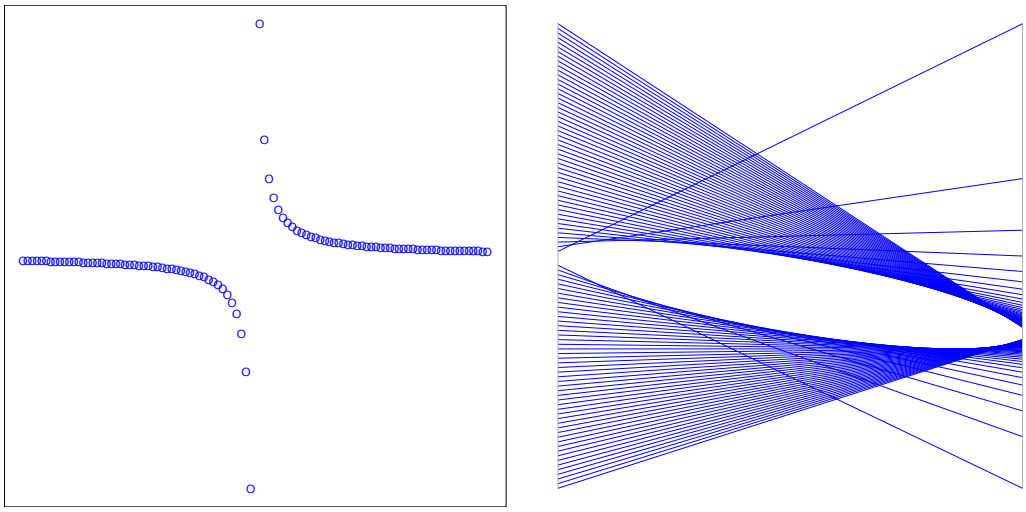

Figure 2.6 Figure illustrant la dualité entre $2D$ et coordonnées parallèles pour la fonction $f : x \to \frac{1}{x}$ sur l'intervalle $[-1, 1]$.

### 2.2.3 Réordonnancement des variables

Les coordonnées parallèles permettent de représenter des données multidimensionnelles. En théorie, il n'y a pas de limite de nombre de variables ou d'observations à représenter. Toutefois, lorsque le nombre de données devient très élevé qu'un grand nombre de lignes se superposent, les graphes en coordonnées parallèles peuvent devenir trop denses et difficiles à analyser. La figure 2.8 montre un exemple de données ou la recherche de modèles et patterns n'est plus évidente.

Le graphique tel qu'il est représenté dans la figure de gauche ne permet pas de montrer que la dépendance entre $x4$ et $x3$. Cependant, un réarrangement manuel des variables permet de ressortir une deuxième dépendance entre les variables $x1$ et $x4$ en les plaçant en adjacence. Ces relations sont détectées à l'aide des formes claires qui apparaissent entre les variables voisines, ainsi que par les lignes représentatives des observations entre les paires de variables se comportant de la même façon pour des observations dans la même zone. Par exemple, entre les variables $x1$ et $x4$, toutes les observations sont modélisées par des lignes parallèles. Plusieurs techniques sont proposées pour améliorer l'exploration visuelle des données en coordonnées parallèles. Ces techniques visent à ordonner les variables de façon à accentuer les dépendances entre les variables et la structure globale des données. Ces techniques sont également utilisées pour réduire le désordre dans les graphiques. L'objectif est, ainsi, l'amélioration de la détection des segments. La figure 2.9 montre comment le réarrangement améliore la visualisation des segments de données.

L'ordre des variables a un impact visible sur la perception des formes entre les variables

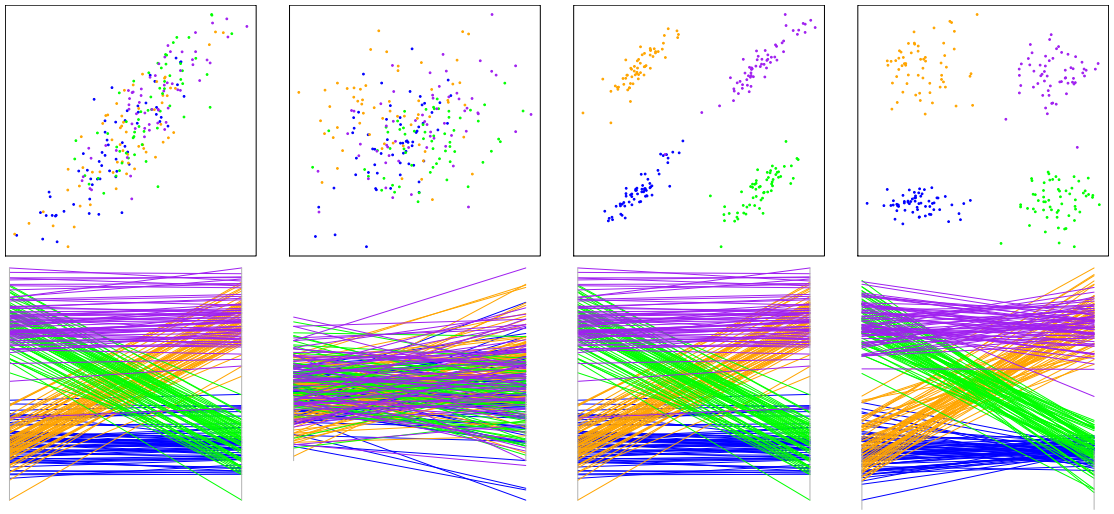

Figure 2.7 Figure illustrant la perception des données séparables ou/ et corrélées en coordonnées Cartésiennes et en coordonnées parallèles..

et sur la visualisation des dépendances entre les variables et des segments de données. En effet, Peng *et al.* (2004) confirme que les coordonnées parallèles permettent de visualiser les relations entre les variables adjacentes, mais ne révèlent pas les variables non adjacentes, d'où, l'importance du choix de l'ordre de variables. Le réarrangement des variables permet de montrer la dépendance entre les données, améliore l'exploration visuelle des données et facilite la compréhension et l'analyse. Les logiciels interactifs permettent d'interchanger les axes parallèles manuellement. L'utilisateur peut, ainsi, choisir l'ordre qu'il souhaite voir et observer les structures qui apparaissent. Lorsque l'utilisateur choisit l'ordre manuellement, il peut omettre ou oublier certaines relations importantes. L'essai de toutes les permutations possibles peut être long et fastidieux. Pour rendre la tache de réarrangement plus simple, plusieurs auteurs ont proposé des techniques de réarrangement automatique des variables.

Ces techniques visent à placer chaque variable dans le voisinage des autres de façon à maximiser un critère défini dans la mesure du possible, car une variable a, au maximum, deux variables voisines. Ankerst *et al.* (1998) propose une technique pour minimiser les dissimilarités ou les dissimilarités partielles (pour une partie des données) entre deux variables adjacentes. Il mesure les dissimilarités en utilisant les distances Euclidiennes. Avant de calculer les distances Euclidiennes, les variables sont standardisées. Cette métrique n'est pas fiable avec tous les types de dépendances, y compris les variables qui sont liées paraboliquement. (Peng *et al.*, 2004) propose une seconde technique pour réordonner les variables. Cette technique vise à réordonner les variables de façon à minimiser les observations aberrantes entre les paires de variables voisines. Une observation est considérée comme aberrante si elle ne possède pas un minimum d'observations voisines. Les variables sont considérées voisines si

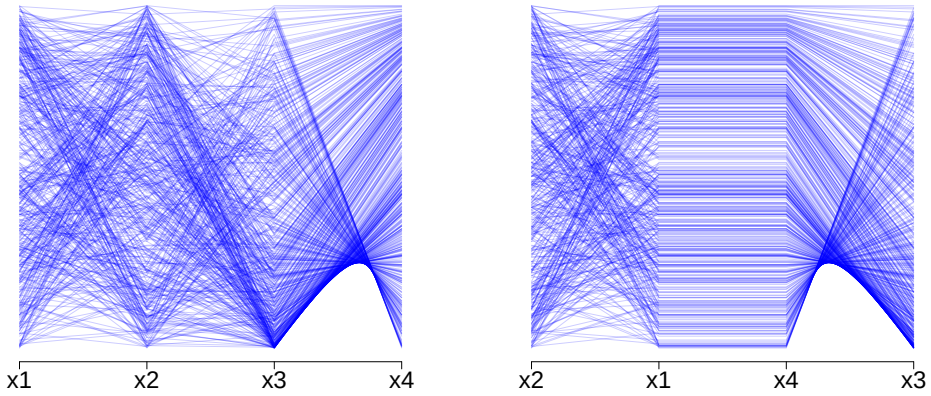


Figure 2.8 Exemple de données avant (figure de gauche) et après réarrangement manuel (figure de droite). Le réarrangement est réalisé dans l'objectif visualiser les dépendances entre les variables.

elles sont distantes de moins d'un certain seuil prédéfini. Le problème de cette méthode est qu'elle est très sensible au choix du seuil et au choix de la taille du voisinage minimum. Deux variables aberrantes peuvent être omises si elles sont voisines.

Johansson et Johansson (2009) étudient différentes métriques qui peuvent être employées pour réordonner, automatiquement, les variables. Ces métriques incluent la maximisation des corrélations, l'amélioration de la qualité de détection de segments et l'amélioration de la détection d'observations aberrantes.

Ferdosi et Roerdink (2011) introduisent une approche utilisant le concept de regroupement en sous-espace et le classement pour le réarrangement des dimensions. Ils classent les sous-espaces selon certains critères de qualité tels que la corrélation de Pearson et la détection des valeurs aberrantes. Ensuite, ils réordonnent les variables. Dasgupta et Kosara (2010) évaluent une technique basée sur plusieurs critères tels que le nombre de lignes qui se croisent, la minimisation des angles d'intersection, la maximisation du parallélisme et l'information de convergence divergence. L'utilisateur doit choisir lui-même le poids à associer à chacune de ces métriques pour trouver l'ordre optimal de réordonnacement des variables. Pour résoudre le problème d'optimisation, ils suggèrent l'algorithme Branch and bound. Leurs techniques testées en utilisant la base de données de la qualité de vin blanc montrent que l'ordre obtenu est différent selon les poids associés aux critères discutés. Un autre algorithme est indépendamment développé en utilisant l'algorithme génétique pour souligner les attributs les plus importants et permettre la détection des irrégularités en utilisant la corrélation de Pearson (Boogaerts *et al.*, 2012). Lu *et al.* (2016) appliquent la décomposition en valeurs singulières

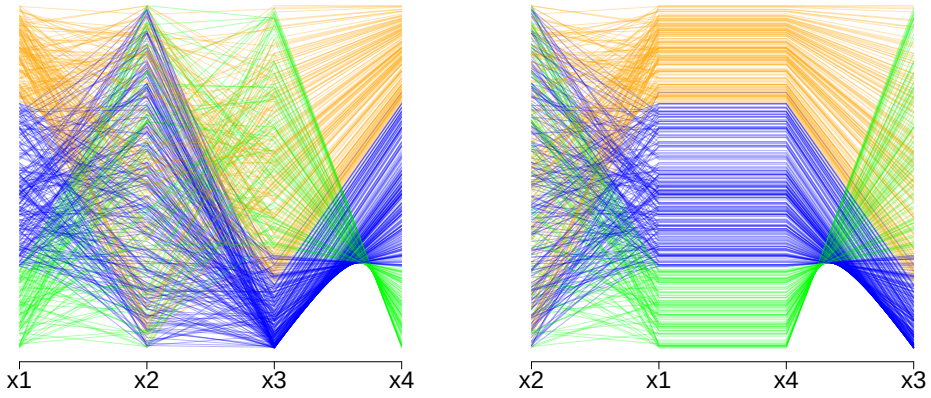


Figure 2.9 Exemple de données avant (figure de gauche) et après réarrangement manuel (figure de droite). Le réarrangement est réalisé dans l'objectif d'améliorer la séparation visuelle des données.

pour sélectionner les attributs qui ont le plus de contribution, ensuite, en fonction des corrélations non linéaires, ils ordonnent les axes. Finalement, PairViz est un pacquage *R* qui produit un ordonnancement d'objets statistiques pour des propos de visualisation. Le problème d'ordonnancement des variables se modélise par un problème de construction d'arêtes d'un graphe. PairViz implémente divers algorithmes de parcours de graphes basés sur les tours Eulériens et les décompositions Hamiltoniennes (Hurley et Oldford, 2011). Cette méthode peut être appliquée à différentes techniques de visualisation de données multidimensionnelles incluant les coordonnées parallèles. Van Long et Linsen (2016) proposent une méthode qui fournit un ordonnancement automatique des axes de coordonnées parallèles. L'idée principale de leur méthode est de trouver une solution sous optimale pour les permutations basées sur les similarités des dimensions. Les 2 métriques de similarité utilisées sont la corrélation de Pearson pour les données non classées et la consistance de distances de classe pour les données classées. Cette section présente une revue des algorithmes de réarrangement des variables pour optimiser la visualisation des coordonnées parallèles. L'objectif de ces techniques est de soutenir l'exploration visuelle des données et de souligner la forme des données. La revue de littérature se poursuit avec les techniques soutenant l'exploration industrielle des données en coordonnées parallèles avec une revue des techniques de segmentation et des graphes de densités en coordonnées parallèles (sous-section 2.2.5).

### 2.2.4 segmentation

Dans cette sous-section, une brève revue de la littérature des techniques de segmentation est présentée. Les algorithmes de segmentation appliqués aux coordonnées parallèles sont discutés. La segmentation vise à trouver des sous-groupes ou segments dans un ensemble de données. En d'autres termes, nous cherchons à séparer les données de manière à ce que les données d'un même segment soient assez similaires et que les données de deux segments différents soient assez dissimilaires. Ainsi, en général, les techniques de segmentation diffèrent par le choix de la définition de deux observations différentes et deux observations similaires. Parmi les techniques de segmentation les plus connues, nous pouvons citer la méthode des $k$ moyennes et la segmentation hiérarchique (James et Tibshirani, 2013). Plusieurs techniques de segmentation sont proposées dans la littérature. Peterson (2002) mentionne les $k$-moyennes, les $k$-medoides, CLARANS, DIANA et BIRCH. Il définit l'algorithme des $k$-moyennes comme un algorithme de segmentation de base qui crée des segments circulaires dans un plan à 2 dimensions et sphériques dans un plan à 3 dimensions. Les segments sont créés autour des centroïdes. Une explication plus détaillée de la méthode des $k$ moyennes est présentée dans le chapitre 4. Le principe des $k$-medoides est très similaire aux $k$-moyennes avec l'exception qu'au lieu de créer des segments autour des centroïdes, l'algorithme du $k$-medoides associe un point au groupe en se basant sur le point le proche. La segmentation hiérarchique est une technique largement utilisée. L'idée derrière cette segmentation est de construire un arbre binaire des données qui fusionne les données en groupes de données similaires (Sharma et Kaur, 2013). Un ensemble de données est réparti en se basant sur certaines mesures de dissimilarité (différence) prédéfinies. La visualisation de cet arbre fournit un résumé de la démarche de segmentation. Ceci est utile pour l'explication ou l'interprétation de la répartition des données. Pour plus de détails sur les techniques de segmentation, Jain et Flynn (1999) est une source assez complète.
La segmentation est fortement implémentée dans les graphes en coordonnées parallèles. Elle permet d'améliorer l'exploration visuelle des données, facilite la détection de structure et réduit le désordre dans les graphes (Johansson et Forsell, 2016).

Différents algorithmes de regroupement sont appliqués aux coordonnées parallèles. La segmentation est implémentée dans plusieurs outils interactifs de coordonnées parallèles tels que XMDVTool et XDat. Les $k$ moyennes sont également utilisées par Fua *et al.* (1999). Fua *et al.* (1999) utilisent la segmentation hiérarchique et visualisent les segments par des bandes d'opacité variables diminuant du centre vers les arêtes. Johansson *et al.* (2005) proposent regrouper les données en segments et les représenter à l'aide de différentes textures afin de régler un problème de coordonnées parallèles, i.e, comment visualiser un nombre élevé de

données sans cacher la structure générale des données. Berthold et Hall (2003) suggèrent également visualiser les données en les séparant en différents segments flous (fuzzy clusters). Dans ce cas, les différents groupes de données sont visualisés de façon à montrer le degré d'appartenance à ces groupes.

### 2.2.5 Graphes de densité en coordonnées parallèles.

Les graphes de densité visualisent les fonctions de distribution des données d'intérêt. Ils permettent de visualiser une fonction lisse continue à la place d'un ensemble de points discrets, permettant ainsi de distinguer les zones plus probables des zones les moins probables. Les graphes de densité visualisent, autrement, les fonctions de densité d'une variable ou d'un vecteur. Soit $X$ une variable aléatoire et $F_X$ sa fonction de répartition, s'il existe une fonction $f_X$ définie et positive sur un intervalle $\Omega$ de $\mathbb{R}$ telle que :

$$\forall x \in \Omega, F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

Cette fonction $f_X$ est appelée la fonction de densité de la variable $X$ et vérifie :

$$\int_{-\infty}^{+\infty} f_X(u)du = 1.$$

La fonction de densité permet de calculer la probabilité d'appartenance de $X$ à un ensemble de données. Les fonctions de distribution sont généralement méconnues pour les données réelles. Elles doivent être estimées soient par des tests d'ajustement à des fonctions de lois connues comme la loi normale ou la loi uniforme, ou exponentielle. Les fonctions de densité peuvent être estimées par des méthodes non paramétriques telles que les histogrammes ou encore les méthodes d'estimation par noyaux. Cette dernière méthode est utilisée pour la conception des cartes, elle est expliquée dans le chapitre 4. Les coordonnées parallèles à densité sont similaires aux graphes de densité bidimensionnelle. Elles représentent une fonction de densité continue des données d'intérêt. De nombreux auteurs ont étudié cet axe de recherche. En fait, les coordonnées parallèles continues visualisent la densité des lignes au lieu de visualiser des lignes discrètes. Artero et al. (cité dans Zhou *et al.* (2008)) proposent des modèles de densité pour traiter le problème d'encombrement et de désordre pendant que les données sont visualisées. Ils suggèrent de filtrer les informations en coordonnées parallèles en se basant sur des informations de fréquence ou de densité calculées à partir de l'ensemble de données. Heinrich et Weiskopf (2009) et Heinrich et Weiskopf (2013) proposent 2 modèles pour la détermination de la densité. Le premier modèle est basé sur le binning et le second est un modèle de densité de lignes. Le binning consiste à estimer la fonction de densité tel que

les données sont regroupées dans les classes d'un histogramme. Pour le second modèle, la fonction de densité est estimée par une fonction de densité Gaussienne.

Les coordonnées parallèles continues, décrites par Heinrich et Weiskopf (2013), transforment les densités d'un nuage de points en densité de coordonnées parallèles bidimensionnelles (c'est-à-dire avec 2 axes) exploitant la dualité point-ligne entre les coordonnées Cartésiennes et les coordonnées parallèles. En fait, la densité de chaque paire de variables est estimée en $2D$ par une loi normale puis transformée en se basant sur les équations de dualité expliquées dans la sous-section 2.2.2. Ensuite, chaque densité est associée à une couleur ou et une transparence bien définie. Heinrich et Weiskopf (2013) proposent des méthodes géométriques qui permettent de visualiser les densités implicites basées sur la proximité des objets géométriques.

Au lieu de représenter chaque point d'observation par une ligne, Palmas *et al.* (2014) suggèrent de remplacer un ensemble de lignes par une bande polygonale. Cela rend le temps de transformation indépendant du nombre d'observations. De cette façon, ils ont essayé de rendre une méthode réactive même pour de très grandes bases de données. Palmas *et al.* (2014) introduisent cette technique basée sur la segmentation et le regroupement pour réduire le désordre et rendre un graphique plus instructif. Cette méthode segmente les données dans chaque dimension de façon indépendante. Ces segments représentés sur chaque axe et chaque ligne (observations) sont groupés pour former des bandes polygonales entre les segments voisins.

Chen *et al.* (2016) présentent une méthode de visualisation de coordonnées parallèles anisotropes. Cette méthode peut améliorer l'efficacité de l'analyse visuelle des données multidimensionnelles. Il s'agit de combiner les coordonnées parallèles avec les caractéristiques de distribution de probabilité pour en conclure les coordonnées parallèles anisotropes. Les données de chaque dimension sont d'abord divisées en plusieurs petits segments. La fréquence des données dans chaque segment doit être déterminée et pourrait être considérée comme la caractéristique de distribution de la dimension correspondante. La distribution des données dans chaque dimension est exprimée par une forme comme un histogramme. L'occupation de chaque segment sur chaque axe de coordonnées est ajusté en fonction de la distribution de la dimension correspondante. En conclusion, plusieurs recherches sont proposées pour concevoir des graphes de densité. La plupart définit les densités ou bien sur chaque dimension séparément ou bien au niveau des paires de variables adjacentes. Notre objectif étant de définir une sorte de chemin pour une observation fonctionnelle, de conserver la forme des données et de visualiser la densité de l'ensemble des observations, nous proposons une nouvelle technique basée sur la dualité entre le plan Cartésien et les coordonnées parallèles et sur l'estimation

de la densité par noyaux.

## 2.3 Conclusion

La revue de littérature regroupe plusieurs concepts dont certains sont liés à la problématique et à l'objectif de travail et d'autres sont reliés à la méthodologie du travail. Dans la revue de littérature, nous revoyons quelques outils multidimensionnels de maitrise de processus et de contrôle qualité. Ceci inclut les cartes de contrôle et les algorithmes d'apprentissage statistique, ainsi qu'une revue des outils de contrôle à support visuel. Cette revue de littérature montre que les outils de contrôle multidimensionnels constituent un champs d'intérêt pour les travaux de recherche malgré certains problèmes, particulièrement dans le diagnostic des défauts et dérives détectés. Ensuite, les coordonnées parallèles qui constituent un type de graphes multidimensionnels, sont introduites. La revue regroupe certains concepts qui lui sont liés tels que la dualité entre le plan Cartésien et les coordonnées parallèles et l'arrangement des variables. Bien qu'actuellement, ces 2 concepts ne sont pas très liés dans la littérature, une sorte de connexion commence à ressortir à travers certaines nouvelles recherches. En effet, certains travaux cherchent à exploiter les graphes multidimensionnels pour proposer des cartes de contrôle multidimensionnelles qui est aussi l'objectif de ce travail. La compréhension des différents concepts présentés dans la revue de littérature permet, également, de comprendre la méthodologie de conception des cartes proposées dans cette thèse expliquée dans le chapitre 3 et le chapitre 4.

# CHAPITRE 3 RÉARRANGEMENT DES VARIABLES

Dans le chapitre 2, l'importance de l'arrangement des attributs en coordonnées parallèles est démontrée. À travers la revue de littérature, plusieurs algorithmes sont proposés. Le but de ces algorithmes est d'améliorer l'exploration visuelle des données. Dans ce chapitre, nous proposons un cadre générique d'arrangement des attributs d'une base de données. Ce cadre s'adapte à l'objectif d'arrangement étudié, i.e, la séparation des données, la détection des données aberrantes et peut couvrir d'autres objectifs. L'arrangement des variables se fait par rapport à une mesure générique appelée information générale définie par rapport à certaines distributions bivariées. Les fonctions de distribution et les métriques sont définies par rapport à l'objectif de réordonnacement.

## 3.1 Information générale

Dans cette section, l'information générale est définie entre chaque paire de variables. Dans la section 3.2, la démarche utilisant l'information générale pour réordonner les attributs est présentée.
Soient $x_1$ et $x_2$ 2 attributs. On définit deux mesures de probabilité hypothétiques sur un même sigma algèbre $\mathcal{F}$. Autrement, nous définissons 2 mesures de probabilité $F$ et $H$ définies sur deux espaces de probabilité $\Big(\Omega, \mathcal{F}, F\Big)$, and $\Big(\Omega, \mathcal{F}, H\Big)$. $F(x_1, x_2)$ et $H(x_1, x_2)$ représentent deux mesures de probabilité différentes pour $x_1$ et $x_2$. Ceci veut dire,

$$\exists (x_1, x_2) \in \mathbb{R}^2 \text{ tel que } F(x_1, x_2) \neq H(x_1, x_2).$$

L'information générale est définie comme :

$$GI(x_1, x_2) = \frac{1}{G''(1)} \int_{x_1} \int_{x_2} G\left\{ \frac{dF(x_1, x_2)}{dH(x_1, x_2)} \right\} dH(x_1, x_2), \tag{3.1}$$

avec $\frac{dF(x_1,x_2)}{dH(x_1,x_2)}$ est la dérivée de Radon-Nikodym et $G(.)$ est une fonction continue. La dérivée seconde de $G()$ au point 1, $G''(1)$ est utilisée pour ajuster l'échelle. Le critère 3.1 est étroitement lié à la divergence Kullback-Leibler, à l'entropie croisée et l'entropie jointe. Kullback Liebler est une mesure de dissimilarité entre deux lois de probabilité, elle mesure la divergence entre 2 mesures de probabilité définies pour une même variable.

Le choix de $F$ par rapport à $H$ définit le concept de mesure et est relié à l'objectif d'arrangement d'attributs. Le choix $G(.)$ définit la statistique de mesure. Un choix commun de

$F$ et $H$ est, respectivement, la probabilité jointe et le produit de probabilités marginales. Ces fonctions de distribution sont choisies lorsque l'intérêt du réarrangement porte sur la dépendance des attributs. $GI$ devient équivalente à la corrélation de Pearson si $F(x_1, x_2)$ est une Gaussienne bivariée. Si l'objectif du réordonnacement est la séparation des données, $F$ est choisie comme un mélange de Gaussiennes et $H$ est une seule Gaussienne bivariée. Le critère qui modélise le nombre de valeurs aberrantes est obtenu en associant une distribution à queue à $F$ comme une distribution de Student et à $H$ une distribution centrée symétrique comme une loi Gaussienne. Le rapport $\frac{F}{H}$ modélise le rapport entre le critère d'intérêt et le critère opposé, c'est à dire, la séparation par rapport à la non-séparation, ou la dépendance par rapport à l'indépendance, etc.

Un choix commun de $G(.)$ est $G(u) = u\log(u)$ qui donne l'information mutuelle de $F$ par rapport à $H$, si $F(x_1, x_2)$ est la probabilité jointe de $(x_1, x_2)$ et $H(x_1, x_2)$ est le produit de probabilités marginales. Nous suggérons certains critères pour $G(.)$.

- $G(.)$ est une fonction lisse univariée.
- $G(.)$ s'annule en 1, c'est à dire $G(1) = 0$ ce critère garantit que si le critère d'intérêt est égal à son opposé, l'information est nulle. Par exemple, si la mesure de dépendance $F(x1, x2)$ est égale à la mesure d'indépendance $H(x_1, x_2)$, alors l'information générale est nulle.
- Sa dérivée première est lisse en 1, c'est à dire, $|G''(u)|$ est bornée dans un petit intervalle de $u$, $u \in (1-\epsilon, 1+\epsilon)$. Cette condition garantit un comportement asymptotique de $GI$ lorsque le nombre d'observations augmente.

Une fois, $F$ et $H$ choisies, il convient de choisir $G$. Plusieurs statistiques de contingence connues sont dérivées par la modification de $G(.)$. En effet,

- $G(u) = 2u\log u$ donne le ratio de log de vraisemblance (Friedman *et al.*, 2001),
- $G(u) = (u-1)^2$ donne la statistique Khi deux de Pearson (Friedman *et al.*, 2001),
- $G(u) = u(1-1/\sqrt{u})$ donne la statistique de Freeman-Tukey (Freeman et Tukey, 1950),
- $G(u) = (1-u)^2/u$ donne la statistique de Neyman (Cressie et Read, 1984),
- $G(u) = u(\sqrt[3]{u^2} - 1)$ donne la statistique de Cressie-Read (Cressie et Read, 1984),
- $G(u) = u\log u$ donne l'information mutuelle (Friedman *et al.*, 2001).

L'information mutuelle est un critère largement utilisé dans la revue de littérature pour ordonner les variables en coordonnées parallèles. Le choix de $F$ et $H$ est primordial, car il définit l'objectif d'arrangement des variables. Plusieurs concepts, autres que ceux déjà mentionnés, peuvent être quantifiés à travers le critère 3.1 comme la dispersion, l'asymétrie, la puissance de prédiction et la multicolinéarité. Le réodonnacement est fait en se basant sur ces concepts pour une meilleure exploration visuelle des données. Le choix de $G(.)$ définit la statistique, mais ne doit pas affecter l'objectif d'arrangement.

Cette section explique la méthodologie de détermination des mesures et critères entre deux variables. Dans les sections suivantes, nous expliquons la démarche suivie pour obtenir un ordre optimal ou proche de l'optimal par rapport à l'objectif de visualisation et nous présentons 2 cas d'application ; l'arrangement des variables pour objectif de dépendance entre les variables et l'arrangement pour objectif de séparation des données.

## 3.2 Optimisation de l'ordre

Soit une base de données contenant $p$ attributs. Le nombre total de permutations possibles des axes en coordonnées parallèles est égal à $p!$. Ainsi, il est impossible de vérifier tous les ordres possibles visuellement. Donc, la recherche automatique d'un ordre optimal vient de soi, particulièrement lorsque les données sont de haute dimension. Une matrice d'information générale $W$ est calculée pour toutes les paires des attributs telle que $W_{ij}$ est égal à $W_{ij} = GI(x_i, x_j)$. Le problème de recherche d'un ordre optimal des attributs consiste à trouver une matrice de voisinage $A = [a_{ij}]$ qui maximise l'information générale totale. Le problème revient au problème d'optimisation suivant :

$$\begin{aligned}
\hat{\mathbf{A}} &= \text{argmax } \|\mathbf{A} \odot \mathbf{W}\| && (3.2)\\
\text{s.t.} & \\
a_{ij} = 0 \quad &\text{or} \quad a_{ij} = 1 && (3.3)\\
\mathbf{a}_i^\top \mathbf{1} &= \mathbf{1}^\top \mathbf{a}_j = 2 && (3.4)\\
a_{ij} &= a_{ji}, && (3.5)\\
\|\mathbf{A}\| &\leq 2q && (3.6)
\end{aligned}$$

ou $\mathbf{a}_i^\top$ est la $i$ème ligne de la matrice $\mathbf{A}$, $\mathbf{a}_j$ est la $j$ème colonne de $\mathbf{A}$, $\odot$ est le produit Hadamart et $\|\mathbf{A}\| = \sum_i \sum_j |a_{ij}|$ est la norme $L_1$ de Frobenius.

La fonction objective $\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} w_{ij}$ dans (3.2) traduit la maximisation de la somme d'information générale entre les attributs adjacents. La contrainte (3.3) impose que la variable de voisinage soit binaire et indique ainsi si 2 attributs sont voisins, c'est à dire, $a_{ij} = 1$ ou pas et $a_{ij} = 1$. La contrainte (3.4) assure qu'une coordonnée est dans le voisinage de 2 autres exactement. La contrainte (3.5) impose la symétrie de la matrice de voisinage, ce qui veut dire que si $x_i$ est dans le voisinage de $x_j$ alors $x_j$ est aussi dans le voisinage de $x_i$. La contrainte (3.6), pour $q < p$, sélectionne uniquement $q$ parmi $p$ attributs pour la visualisation. Cette contrainte est particulièrement importante lorsque le nombre total d'attributs dépasse le nombre maximal que nous pouvons visualiser sur un écran (environ 50 attributs).

Si ce n'est pas le cas que tous les attributs doivent être visualisés, nous imposons $q = p$. Les solveurs d'optimisation tels que CPLEX peuvent résoudre ce type de problème en enlevant la contrainte (3.6). L'ajout de cette contrainte fait en sorte que l'algorithme ne converge pas toujours. Le résultat du problème d'optimisation est une matrice d'adjacence, ce qui permet d'avoir un graphe cyclique et non pas une liste. Pour le transformer en une liste, une relation doit être supprimée. Nous choisissons de supprimer la relation dans le graphe cyclique dont la valeur de l'information générale est la plus basse dans le cycle. Ensuite, si nous souhaitons sélectionner exactement $p$ variables, ceci demandera un autre traitement pour ne garder qu'une liste de $q$ attributs maximisant le critère. Pour résoudre ce problème, nous proposons un algorithme plus rapide en cherchant une solution sous optimale de la fonction objective en utilisant l'algorithme glouton. La première paire d'attributs est celle qui maximise l'information générale dans la première itération, c'est à dire,

$$\begin{aligned} (\hat{x}_1, \hat{x}_2) &= \text{argmax } GI(x_i, x_j) \\ 1 \leq i \leq p-1 & \qquad i+1 \leq j \leq p. \end{aligned} \tag{3.7}$$

Les 2 premières coordonnées peuvent être soit $(\hat{x}_1, \hat{x}_2)$ soit $(\hat{x}_2, \hat{x}_1)$. La suite est faite avec les 2 cas possibles, offrant ainsi 2 séquences. La $j$ème, $j = 3, \ldots, q$ coordonnée est choisie telle que

$$\begin{aligned} \hat{x}_j &= \text{argmax } GI(\hat{x}_{j-1}, x_i), \\ i &\in \{1, \ldots, p\} \backslash \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{j-1}\}. \end{aligned} \tag{3.8}$$

$\{1, \ldots, p\} \backslash \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{j-1}\}$ est l'ensemble de variables duquel sont éliminées les variables déjà placées dans la liste. Quand les 2 séquences sont obtenues, la somme d'information totale pour chaque séquence est calculée. La séquence choisie est celle dont l'information totale est maximale. L'exécution de l'algorithme glouton est de complexité $O(p^2)$, la première itération est la plus longue, étant donné que le choix se fait parmi toutes les paires possibles, donc, $\frac{(p^2-p)}{2}$. Si le premier attribut est fixé manuellement, l'algorithme devient plus rapide de complexité $O(pq)$, car le chois du premier attribut se limite, ainsi, à $p$ possibilités. Le choix des coordonnées suivantes se base sur le même principe présenté dans 3.8. Dans la suite, nous présentons 2 cas d'application du cadre général pour des objectifs prédéfinis.

## 3.3 Application du cadre général pour objectif de dépendance

Lorsque l'objectif de réarrangement est fixé, la première étape consiste à fixer les deux mesures de probabilité $F$ et $H$ qui reflètent cet objectif. Pour ordonner les attributs de façon à

maximiser les dépendances, $F$ et $H$ sont choisies comme la probabilité jointe et le produit de probabilités marginales. Très souvent, étant donné que les lois exactes ne sont pas connues,ces deux mesures de probabilité sont estimées en discrétisant les variables. Ainsi, les fonctions de densité correspondant à $F$ et $H$ sont approximées par $p(x_1, x_2)$ et $p(x_1)p(x_2)$ comme pour trouver les classes et les probabilités correspondantes d'un histogramme. Le nombre de classes choisi pour chaque variable est trouvé avec la règle de Sturges. Le nombre de classes est $k = \log_2 n + 1$, ou $n$ le nombre total d'observations. Les fonctions de probabilités marginales de $x_1$ et $x_2$ sont exprimées comme suit :

$$p(x_1) = \frac{\#\{C_i^{x_1}\}}{n};$$

$$p(x_2) = \frac{\#\{C_i^{x_2}\}}{n}.$$

La probabilité jointe est déterminée par l'équation suivante :

$$p(x_1, x_2) = \frac{\#\{C_i^{x_1}, C_j^{x_2}\}}{n}$$

$\#\{C_i^{x_1}\}$ est le nombre de points tels que $x_1 \in \{C_i^{x_1}\}$, $\#\{C_i^{x_2}\}$ est le nombre de points tels que $x_2 \in \{C_i^{x_2}\}$, $\#\{C_i^{x_1}, C_j^{x_2}\}$ est le nombre d'observations $(x_{1k}, x_{2k})_{0\leq k\leq n}$, $x_{1k} \in C_i^{x_1}$ et $x_{2k} \in C_j^{x_2}$. $(C_i^{x_1})_{1\leq i\leq k}$ et $(C_j^{x_2})_{1\leq j\leq k}$ représentent respectivement les classes $x_1$ et $x_2$. L'intégrale dans l'expression de l'information générale est approximée par la somme, ce qui veut dire que l'information générale entre 2 attributs s'exprime comme suit :

$$GI(x_1, x_2) = \frac{1}{G''(1)} \sum_{x_1} \sum_{x_2} G\left\{\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right\} p(x_1)p(x_2), \tag{3.9}$$

$G(.)$ peut être choisie parmi les fonctions mentionnées précédemment pour générer les différentes statistiques. La différence dans le choix de $G(.)$ est discutée dans le chapitre 5. La démarche de discrétisation et de calcul de l'information générale est appliquée à chaque paire d'attributs pour obtenir la matrice d'information générale. Ensuite, l'ordre optimal est obtenu en résolvant le problème d'optimisation présenté dans la section 3.2 soit avec CPLEX ou avec l'algorithme Glouton.

## 3.4 Application du cadre général pour objectif de séparation

Si l'objectif d'ordonnancement est l'amélioration de la qualité de détection de segments de données, $F$ est choisie comme un mélange de Gaussiennes et $H$ est choisie comme une Gaus-

sienne unique.

$$dF(x_1, x_2) = f(x_1, x_2) = \sum_c \pi_c g_c(x_1, x_2)$$

$g_c(x_1, x_2)$ est une fonction Gaussienne bivariée, représentant la composante $c$, de fonction de densité :

$$g_c(x_1, x_2) = \frac{1}{2\pi|\mathbf{\Sigma}|}\exp\{\frac{-1}{2}(\mathbf{x} - \mu_k)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_k)\}$$

$|\mathbf{\Sigma}|$ est le déterminant de la matrice de variance-covariance $\mu_k = (\mu_{1k}, \mu_{2k})$ est estimée par la moyenne de chaque segment de données,

$$\mathbf{\Sigma}_k = \begin{bmatrix} \sigma_{1k} & 0 \\ 0 & \sigma_{2k} \end{bmatrix} \tag{3.10}$$

$\sigma_{1k} = \frac{1}{k}\sigma_1$, $\sigma_{2k} = \frac{1}{k}\sigma_2$ $\sigma_1$ est estimée par l'écart type de la variable $x_1$ et $\sigma_2$ est estimée par l'écart type de la variable $x_2$. La covariance entre les variables $x_1$ est $x_2$ est supposée nulle comme nous ne nous intéressons qu'a la séparation des données. $\pi_k$ est la probabilité d'appartenir à un segment de données $C_k$, $\#_k$ est le nombre d'observations dans un segment $C_k$ et $n$ est le nombre d'observations total. $\pi_k$ est estimée à $\pi_k = \frac{\#_k}{n}$. $dH(x_1, x_2) = g(x_1, x_2)$ ou $g(x_1, x_2)$ est une fonction Gaussienne bivariée à covariance nulle. Pour trouver le mélange de Gaussiennes, les paramètres $\mu_k$ et $\sigma_k$ doivent être estimés. Cependant, l'ajustement à un mélange de Gaussiennes peut s'avérer très long et très complexe même avec un petit nombre de segments. Ainsi, une démarche alternative plus rapide est proposée. Cette démarche consiste à appliquer la méthode des $k$-moyennes à la place d'ajuster les données à un mélange de Gaussiennes. Les données sont réparties en $k$ segments. La moyenne de la composante $c$ est estimée par la moyenne du segment $c$. La probabilité associée à la composante $c$, $\pi_c$ est estimée à partir du nombre de points qui appartient à chaque segment $c$. Si le nombre de segments adéquat est complètement méconnu, nous suggérons utiliser un nombre élevé de segments de données et de l'ajuster au besoin lors de la visualisation.

Le choix de $G(.)$ n'est pas le plus important ici. Un choix assez commun est $G(u) = u\log(u)$. L'information générale est calculée entre chaque paire de variables et l'ordre optimal est obtenu avec l'algorithme Glouton ou avec le solveur CPLEX. L'ajustement du mélange de Gaussiennes étant couteux en termes de temps, une méthode alternative pour trouver les paramètres de chacune des Gaussiennes. Cette méthode consiste à regrouper les données avec la méthode $k$ moyennes. Le même cadre peut être aussi utilisée pour réordonner les variables selon un critère qui dépend de la sortie $y$ si le contexte est un contexte de classification ou de régression dans ce cas, le critère est remplacé par un critère conditionnel par rapport à $y$. Dans le cas de critère conditionnel, la même démarche est réalisée, mais avec un critère

conditionnel calculé par rapport à $y$ et intégrée par rapport à une variable de sortie $y$.

## 3.5 Explication des étapes de l'algorithme d'arrangement de variables à travers un cas simple : critère de dépendance

Dans cette section, l'arrangement des variables est expliqué à travers un exemple simple. Soient $x_1$, $x_2$ et $x_3$ trois variables telles que $x_1$ et $x_2$ sont aléatoirement et respectivement entre $[0, 1]$ et $[-2, 1]$ et $x_3 = 2x_1$. Soit $n = 30$, le nombre d'observations générées. Nous choisissons d'ordonner les variables selon le critère de dépendance. $F$ est alors approximée par $p(x_1, x_2)$ et $H$ par $p(x_1)p(x_2)$, de même pour $F(x_1, x_3)$, $H(x_1, x_3)$, $F(x_2, x_3)$ et $H(x_2, x_3)$. $G(u) = u\log(u)$ La première étape consiste à estimer les fonctions de distribution suivantes pour déterminer les valeurs ponctuelles de $F$ et $H$.

$$p(x_1), p(x_2), p(x_3), p(x_1, x_2), p(x_1, x_3), p(x_2, x_3).$$

Comme nous trouvons les fréquences des classes d'un histogramme, la fonction de distribution de la variable $x_1$ est déterminée. Elle est exprimée comme suit :

$$p(x_1) = \frac{6}{30} \quad \text{if} \quad x_1 \in [0, 0.2[ \tag{3.11}$$

$$p(x_1) = \frac{8}{30} \quad \text{if} \quad x_1 \in [0.2, 0.4[ \tag{3.12}$$

$$p(x_1) = \frac{4}{30} \quad \text{if} \quad x_1 \in [0.4, 0.6[ \tag{3.13}$$

$$p(x_1) = \frac{9}{30} \quad \text{if} \quad x_1 \in [0.6, 0.8[ \tag{3.14}$$

$$p(x_1) = \frac{3}{30} \quad \text{if} \quad x_1 \in [0.8, 1[ \tag{3.15}$$

Les fonctions de probabilité de $x_2$ et $x_3$ sont déterminées de la même façon. Sans détailler, la fonction de probabilité de $x_2$, nous donnons 2 valeurs comme exemples :

$$p(x_2) = \frac{5}{30} \quad \text{if} \quad x_2 \in [-2, -1.5[ \tag{3.16}$$

$$p(x_2) = \frac{7}{30} \quad \text{if} \quad x_1 2 \in [-1.5, -1[ \tag{3.17}$$

Pour la probabilité jointe de $(x_1, x_2)$, elle est déterminée par le nombre d'observations qui appartiennent à chacune des $5 \times 5$ classe. Par exemple,

$$p(x_1, x_2) = \frac{1}{30} \quad if \quad (x_1, x_2) \in ([0, 0.2[, [-2, -1.5[) \tag{3.18}$$

$$p(x_1, x_2) = \frac{2}{30} \quad if \quad (x_1, x_2) \in ([0, 0.2[, [-1, -0.5[) \tag{3.19}$$

Ainsi,

$$GI(x_1, x_2) = \frac{1}{G''(1)} \sum_{x_1} \sum_{x_2} G \left\{ \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right\} p(x_1)p(x_2)$$

Or, $G''(u) = \frac{1}{u}$, donc, $G''(1) = 1$.

$$GI(x_1, x_2) = \sum_{x_1} \sum_{x_2} \left\{ \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right\} \log \left\{ \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right\} p(x_1)p(x_2)$$

$$GI(x_1, x_2) = \sum_{x_1} \sum_{x_2} \{p(x_1, x_2)\} \log \left\{ \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right\}$$

$$GI(x_1, x_2) = \frac{1}{30} \log \left\{ \frac{\frac{1}{30}}{\frac{6}{30}\frac{5}{30}} \right\} + \frac{2}{30} \log \left\{ \frac{\frac{2}{30}}{\frac{6}{30}\frac{5}{30}} \right\} + \dots$$

Le calcul est achevé, identiquement, pour calculer $GI(x_1, x_2)$, $GI(x_1, x_3)$ et $GI(x_3, x_2)$. Les valeurs obtenues sont $GI(x_1, x_2) = 0.44 \leq GI(x_2, x_3) = 0.46 \leq GI(x_3, x_1) = 1.59$. Selon l'algorithme de réarrangement basé sur l'algorithme de glouton, 2 ordres sont possibles :

— $(x_1, x_3, x_2)$ avec une somme d'information totale de 2.05.

— $(x_3, x_1, x_2)$ avec une somme d'information totale de 2.03.

L'ordre gardé est, alors celui qui maximise l'information totale, $(x_1, x_3, x_2)$.

## 3.6 Conclusion

Ce chapitre présente un cadre général de réarrangement de variables, mais également de sélection des variables à visualiser. Le cadre présenté s'adapte à différents objectifs de visualisation de données. Dans cette thèse, l'intérêt est particulièrement porté sur 2 applications qui sont la dépendance des attributs et la séparation des données. En effet, ceci s'aligne avec l'objectif de développement des cartes de contrôle permettant de distinguer un défaut d'un point fonctionnel. L'étude des dépendances permet d'améliorer la détection des limites de la best opérating zone (cf. Chapitre 4). La séparation des données permet d'obtenir des segments de données plus distinguable. L'impact du réarrangement sur la visualisation des données est évalué dans le chapitre. Le réarrangement constitue la première du développement des cartes

de contrôle.

# CHAPITRE 4 CARTES DE CONTRÔLE MULTIDIMENSIONNELLES

Dans cette section, la méthodologie de développement des 2 types de cartes de contrôle multidimensionnelles est présentée. Les cartes de contrôle proposées sont visualisées avec les coordonnées parallèles. La première version est basée sur la détection de l'enveloppe de la zone de bon fonctionnement (appelée aussi Best Operating zone) et la segmentation de cette zone. La deuxième version proposée pour les données moins nombreuses, pour lesquelles la détection de l'enveloppe de la BOZ n'est pas possible, est la carte de contrôle basée sur les graphes de densité. Le développement de ces cartes nécessite la disponibilité d'une base de données historiques. Les étapes de conception des cartes sont illustrées par la figure 4.1. Après l'obtention des données historiques, un prétraitement est réalisé dépendamment de la

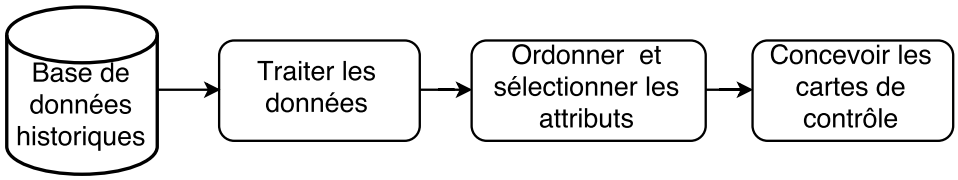


Figure 4.1 Étapes de développement des cartes de contrôle.

base de données. Par exemple, le prétraitement peut être à propos des données manquantes, des variables catégoriques qualitatives qui doivent être transformées en variables catégoriques quantitatives ou encore la suppression des valeurs aberrantes. Les données fonctionnelles sont séparées des données qui représentent les défauts. La première étape consiste à appliquer l'algorithme de réordonnancent de variables. Seules les données fonctionnelles sont utilisées pour ordonner les données. Celles-ci sont réordonnées pour objectifs de dépendance ou séparation. La dépendance permet d'obtenir une enveloppe de BOZ claire et significative. En effet, si les données sont indépendantes, il parait inintéressant de se fier aux limites des relations qui leur relient. L'arrangement des variables selon le critère de séparation des données parait plus adéquat. Il améliore la perception des segments de fonctionnement. L'étape d'arrangement des variables est primordiale, elle est expliquée dans le Chapitre 3.

Pour la version basée sur la BOZ, une fois les données historiques fonctionnelles sont ordonnées, la courbe enveloppe de la BOZ est identifiée par une approche géométrique comme expliquée dans la Section 4.1. Cette enveloppe caractérise le comportement des données fonctionnelles sous contrôle. Ensuite, la BOZ est divisée en différents segments de fonctionnement

plus précis tel qu'expliqué dans la section 4.3. Les nouvelles observations sont visualisées par dessus les limites de la BOZ et les segments de fonctionnement.
Pour la version de densité, les données fonctionnelles ordonnées sont utilisées pour estimer la fonction de densité par une approche d'estimation par noyaux combinée à une approche géométrique utilisée pour la conception des 2 types de cartes. Cette approche consiste à projeter les graphes en coordonnées parallèles dans un plan Cartésien à 2 dimensions (cf. section 4.1).

## 4.1 Projection de points en coordonnées parallèles dans un espace Cartésien à 2 dimensions

Cette section présente la méthode géométrique utilisée pour passer d'un espace en coordonnées parallèles a un plan Cartésien en 2 dimensions. Cette étape constitue le point de départ de la conception des 2 types de cartes proposées. L'approche utilisée est basée sur la dualité entre l'espace Cartésien à 2 dimensions et l'espace en coordonnées parallèles. Entre autres, l'espace en coordonnées parallèles est projeté dans le plan Cartésien. Nous expliquons, d'abord, la projection d'un ensemble d'observations représentées entre 2 coordonnées parallèles. La généralisation est, ensuite, simple. Avant de commencer tout calcul, il est important de standardiser les données afin d'obtenir des enveloppes claires et visibles pour les cartes BOZ et des graphes de densité lisible pour les cartes densité.
Soient $x_1$ et $x_2$, 2 variables adjacentes qui prennent leurs valeurs respectivement dans les intervalles $[0, 1]$ et $[-1000, 0]$. La représentation en coordonnées parallèles sans standardisation implique que la première variable est totalement condensée vers le haut. La distinction des différentes observations devient, alors, impossible. La standardisation des données selon les variables revient à une mise a l'échelle qui dépend des valeurs de chaque variable. Une alternative est de considérer des échelles différentes. Mais comme la projection se réalise dans un même plan Cartésien, l'axe vertical serait le même. Ainsi, pour simplifier la projection expliquée par la suite, toute variable $x_i$ est standardisée comme suit :

$$x_i^s = \frac{x_i - \max\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

ou $\max\{x_i\}$ and $\{x_i\}$ sont le maximum et le minimum de la variable $x_i$. L'impact de la standardisation est illustré par la figure 4.2. Avant la standardisation, la variable $x1$ prend ses valeurs en presqu'un seul point (visuellement), d'où l'importance de la standardisation. La projection est réalisée avec les données standardisées.
Soient $a_1$ et $a_2$ respectivement les axes horizontal et vertical de l'espace Cartésien. Soient $x_1$ et $x_2$, 2 variables, représentées en coordonnées parallèles. Supposons que les axes parallèles sont

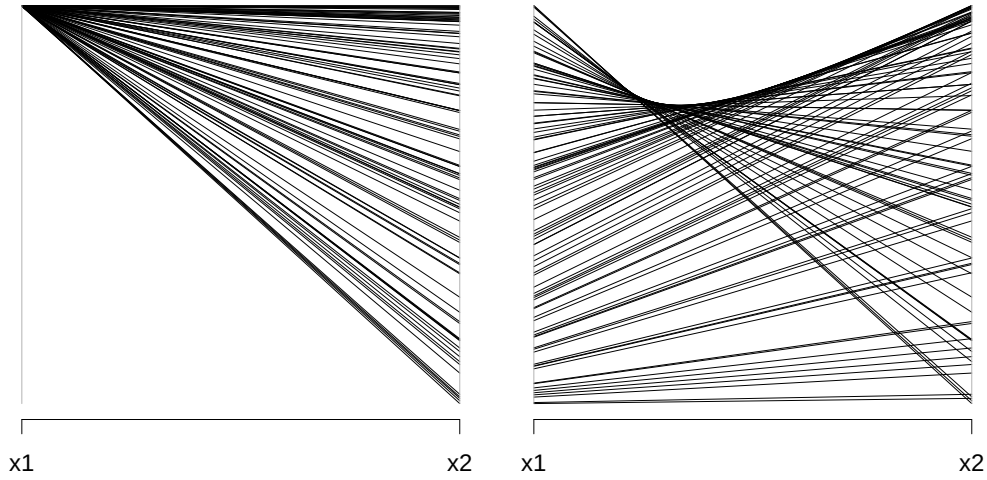


Figure 4.2 Données avant et après standardisation.

distants d'une distance $d$ unités. Les abscisses des axes parallèles $x_1$ et $x_2$ sont respectivement 0 et $d$. L'idée de projection consiste à considérer l'ensemble de lignes qui relient les axes de coordonnées parallèles comme un ensemble de segments de droite dans le plan Cartésien à 2 dimensions.

Dans la figure 4.3, la ligne bleue illustrant l'observation, $(0.87, 0.8)$, dont les extrémités sont visualisées par les points en vert est considérée comme un segment de droite dans le plan Cartésien. Soit $n$ le nombre d'observations et $i \in \{1, \ldots, n\}$. L'équation de la droite qui représente la première observation $(x_{11}, x_{21})$ et qui relie les axes parallèles dans le plan Cartésien est :

$$a_2 = f_1(a_1) = (1 - \frac{a_1}{d})x_{11} + a_1\frac{x_{21}}{d} \tag{4.1}$$

Pour des raisons de simplification, nous supposons que la distance entre les axes parallèles est $d = 1$. Ainsi, l'équation 4.1 devient :

$$a_2 = f_1(a_1) = (1 - a_1)x_{11} + a_1x_{21} \tag{4.2}$$

avec $f_1$ la fonction affine de la droite (en bleu dans la Figure 4.3), $(a_1, a_2) \in [0, 1]^2$ représente le système de coordonnées du plan Cartésien. Dans l'exemple illustré par la figure 4.3, l'équation de la ligne représentant le point $(0.87, 0.8)$ est :

$$a_2 = f_1(a_1) = 0.87(1 - a_1) + 0.8a_1 \tag{4.3}$$

De la même façon, l'équation de la ligne représentant l'observation $i$ entre les axes parallèles

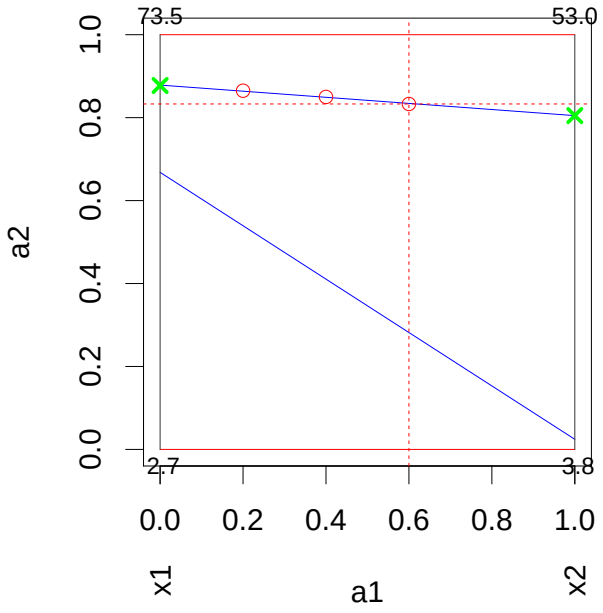


Figure 4.3 Figure expliquant la projection de points représentés en coordonnées parallèles dans un espace Cartésien à 2 dimensions.

$x_1$ et $x_2$, soit $f_i(a_1)$, $a_1 \in [0,1]$), est obtenu pour chaque $i \in \{1,\ldots,n\}$. L'image de l'intervalle $[0,1]$, $f_{i,i\in\{1,\ldots,n\}}([0,1])$ est approximée par l'image de $n_e$ échantillons distants d'un pas $p = \frac{1}{n_e}$ de l'intervalle $[0,1]$. Le résultat donne alors, les ordonnées des points appartenant aux lignes représentant les observations en coordonnées parallèles (points rouges de la Figure 4.3. Ainsi au vecteur de points $A_1 = (a_{11}, a_{12}, \ldots, a_{1n_e})$, correspondrait une matrice de dimension $(n, n_e)$ comportant les images de chaque point du vecteur par les fonctions $f_{i,i\in 1,\ldots,n}$ :

$$\mathbf{Y} = \begin{pmatrix} f_1(a_{11}) & f_1(a_{12}) & \cdots & f_1(a_{1n_e}) \\ f_2(a_{11}) & f_2(a_{12}) & \cdots & f_2(a_{1n_e}) \\ \vdots & \vdots & \ddots & \vdots \\ f_n(a_{11}) & f_n(a_{12}) & \cdots & f_n(a_{1n_e}) \end{pmatrix} \tag{4.4}$$

La figure 4.4 montre un ensemble de données représentées en coordonnées parallèles et le résultat de la projection dans le plan Cartésien. Elle correspond la projection de données bidimensionnelles. Si nous souhaitons projeter des données multidimensionnelles, la démarche expliquée doit être répétée entre chaque paire de variables adjacentes.

Nous passons maintenant aux étapes de conception des 2 types de cartes.

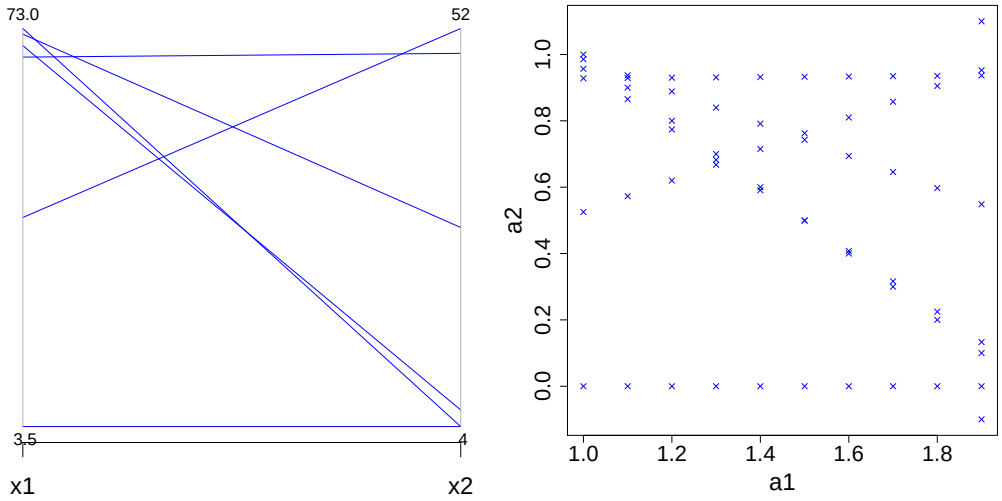


Figure 4.4 Projection de points en coordonnées parallèles dans un plan Cartésien.

## 4.2 Développement des cartes BOZ

Les étapes de développement des cartes BOZ sont résumées dans la figure 4.5. La projection

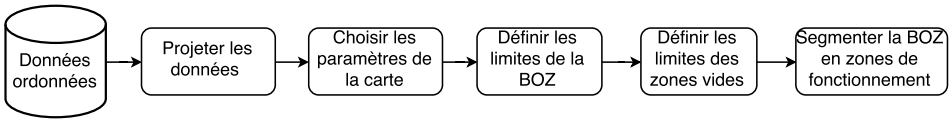


Figure 4.5 Étapes de conception des cartes BOZ.

est expliquée dans la section 4.1. Le choix des paramètres se fait avec la méthode de validation croisée avant la définition des limites de la BOZ et avant sa segmentation. Cependant, pour comprendre la méthode de validation croisée appliquée en particulier a ce contexte, il parait cohérent de présenter les étapes de la zone sous contrôle avant (BOZ et segments de fonctionnement).

### 4.2.1 Identification de l'enveloppe de la Best Operating Zone

Quand les données fonctionnelles sont représentées, nous pouvons apercevoir certaines formes apparaitre au niveau des limites, particulièrement, lorsque les variables adjacentes sont dépendantes, tel qu'expliqué dans la section 2.2.2 et telle que le montre la figure 4.6. La figure 4.6 représente 3 variables visualisées en coordonnées parallèles. Cette figure montre des courbes et des lignes qui délimitent les observations représentées. Une courbe elliptique en haut et un

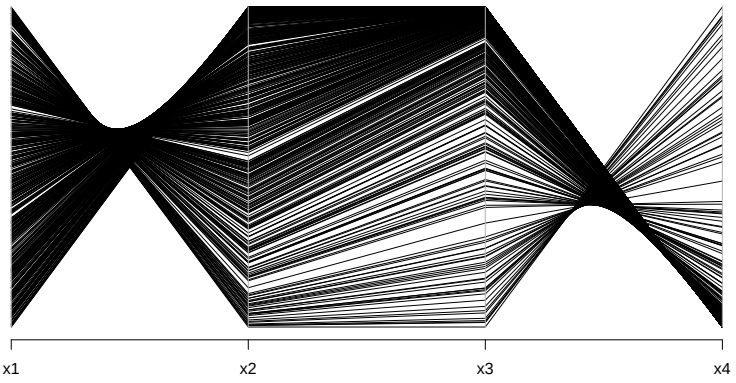


Figure 4.6 Exemple de formes qui apparaissent en coordonnées parallèles.

triangle en bas apparaissent entre les variables $x_1$ et $x_2$. Les points sont limités par 2 lignes horizontales entre les variables $x_2$ et $x_3$. Finalement, une courbe parabolique en bas et un triangle en haut délimitent les points entre $x_3$ et $x_4$.
Pour déterminer les limites de la BOZ, les données standardisées représentées en coordonnées parallèles sont projetées dans le plan Cartésien. Afin de trouver les limites supérieures et inférieures entre les variables adjacentes $x_1$ et $x_2$, le maximum et le minimum sont calculés sur chaque colonne de la matrice de projection (cf. section 4.1). Pour chaque abscisse de l'axe $a_1$, le minimum et le maximum sont retenus. Les coordonnées des points de la courbe enveloppe supérieure, entre 2 axes parallèles, sont :

$$\Big[\frac{j}{n_e}, \max_{i,i\in\{1,\ldots,n\}}(f_i(\frac{j}{n_e}))\Big].$$

Les coordonnées des points de la courbe enveloppe inférieure sont :

$$\Big[\frac{j}{n_e}, \min_{i,i\in\{1,\ldots,n\}}(f_i(\frac{j}{n_e}))\Big],$$

avec $j$ qui varie entre $\{1,\ldots,n_e\}$. La démarche expliquée est appliquée à toutes les paires de variables adjacentes, l'adjacence étant déterminée par l'algorithme d'arrangement de variables. Ceci permet de déterminer les coordonnées des courbes enveloppes multidimensionnelles. Ceci permet d'obtenir une courbe enveloppe supérieure et une courbe enveloppe inférieure tels qu'illustré par la figure 4.7. Pour éviter le surajustement des cartes de contrôle aux données d'apprentissage historiques, les limites supérieures et inférieures sont ajustées de $\epsilon_l$. Les coordonnées des limites inférieures et supérieures sont, alors respectivement :

$$\Big[\frac{k+j}{n_e}, \max_{i,i\in\{1,\ldots,n\}}\{f_i^k(\frac{j}{n_e})\} + \epsilon_l\Big],$$

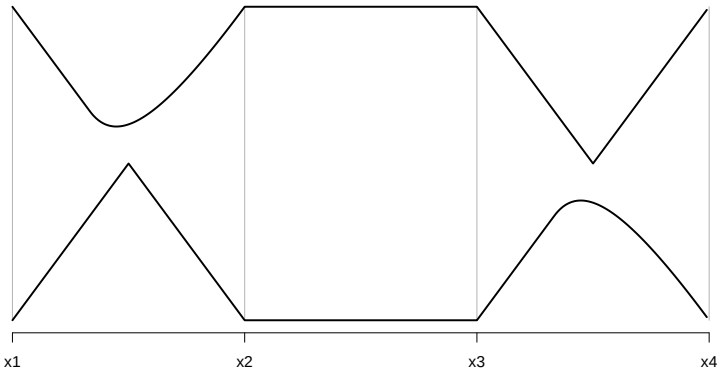


Figure 4.7 Exemple montrant les limites de la BOZ.

$$\Big[\frac{k+j}{n_e}, \min_{i,i\in\{1,\dots,n\}}\{f_i^k(\frac{j}{n_e})\} + \epsilon_l\Big],$$

ou $j \in \{1,\dots,n_e\}$ et $k \in \{1,\dots,p-1\}$. $\max_{i,i\in\{1,\dots,n\}}\{f_i^k(\frac{j}{n_e})\}$ et $\min_{i,i\in\{1,\dots,n\}}\{f_i^k(\frac{j}{n_e})\}$ prennent leurs valeurs dans l'intervalle $[0,1]$, puisque les données sont standardisées. $\epsilon_l$ est choisie à l'aide d'un algorithme de validation croisée expliquée à la fin de la section.
Cependant, si les données ne sont pas définies sur une zone intérieure aux limites supérieure et inférieure, cette zone doit être, également, exclue de la BOZ. Un exemple est donné par la Figure 4.8. La zone vide en forme d'ellipse entourée en rouge est la zone a exclure de la BOZ.

Pour déterminer les limites de la zone en question, chaque colonne de la matrice de projection est triée par ordre croissant, c'est-à-dire les images du point $\frac{j}{n_e}$ sont triées par ordre croissant. Ensuite, la distance $f_i^k(\frac{j}{n_e}) - f_l^k(\frac{j}{n_e})$, représentant la largeur de la zone vide, est calculée pour $j \in \{1,\dots,n_e\}$, $i \in \{1,\dots,n\}$ et $k \in \{1,\dots,p\}$. $f_l^k(\frac{j}{n_e})$ est la plus proche image par $f_l^k$ telle que $f_i^k(\frac{j}{n_e}) \geq f_l^k(\frac{j}{n_e})$.

Si cette distance dépasse un certain seuil alors les images du point $a_{1i}$ par la fonction par $f_j$ et par $f_{j+1}$ sont sauvegardées et nous passons a l'abscisse suivante $a_{1(i+1)}$ et nous reprenons la même procédure. Le seuil à partir duquel nous considérons une zone vide comme une zone hors contrôle est également fixé par la méthode de validation croisée. Ces étapes sont également répétées pour chaque couple de variables adjacentes telles que pour les limites supérieures et inférieures.

Après la caractérisation de la BOZ par l'identification des courbes enveloppes et des limites des zones vides, des zones de fonctionnement plus précises sont identifiées a l'aide d'un algorithme de segmentation expliqué dans la section 4.3.

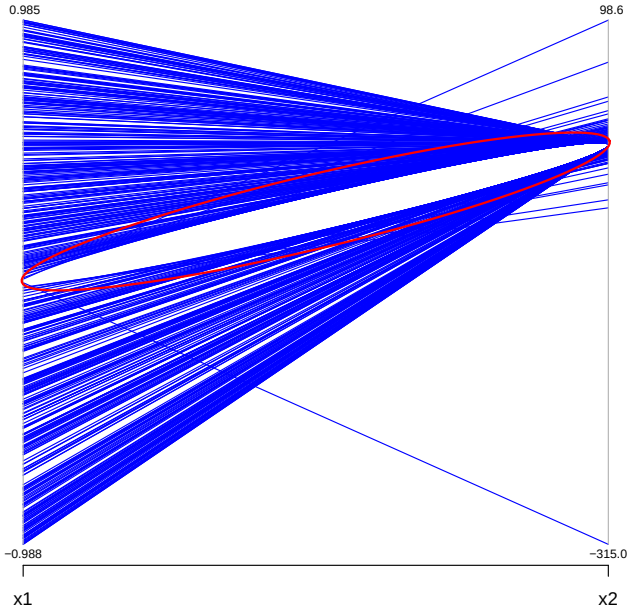


Figure 4.8 Exemple d'une zone intérieure vide qui ne fait pas partie de la BOZ.

## 4.3 Segmentation

La segmentation est appliquée dans l'objectif de mieux définir les zones fonctionnelles. En effet, basé sur les limites de la BOZ, nous classons avec certitude les points en dehors de la BOZ comme des défauts. Cependant, les points à l'intérieur de la BOZ ont une plus grande probabilité d'être fonctionnels. Toutefois, la probabilité que ces points représentent des défauts n'est pas nulle. Pour réduire la probabilité de non-détection de défauts, nous précisons plus la BOZ en qualifiant le comportement des observations fonctionnelles, par la segmentation.

La technique appliquée est la méthode des $k$ moyennes. Elle est choisie, car elle est très souvent appliquée, elle est rapide et elle donne d'assez bons résultats. Si la méthode des $k$ moyennes ne donne pas les résultats souhaités au niveau visuel, la méthode des $k$ médoides peut être utiliser comme alternative. Cette dernière méthode est particulièrement robuste lors de la présence de valeurs aberrantes. En fonction de la base de données, les données sont standardisées ou pas. Lorsque l'ordre de grandeur des données est très différent, la standardisation devient nécessaire.

$$x^s{}_i = \frac{x_i - \max\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

ou $\max\{x_i\}$ et $\min\{x_i\}$ sont le maximum et le minimum de la variable $x_i$.

La technique des $k$ moyennes est une technique qui vise à séparer les données en $k$ segments non superposés. Les segments sont construits de façon à ce que la somme des variations inter-segments est la plus petite possible et à ce que la somme des variations entre les différents segments est maximale. Les variations sont souvent mesurées à l'aide des distances, soient par exemple les distances Euclidiennes. Dans la méthode des $k$ moyennes, les données sont regroupées autour des moyennes de segments. En fait, $k$ points sont au départ choisis aléatoirement. Ces points sont considérés comme centres des segments. Les autres points de données sont ensuite associés au $k$ segments selon leur distance par rapport aux centres. Un point est associé au segment dont la distance avec son centre est minimale. Une fois, toutes les observations associées à un segment, les nouveaux centres de segments sont déterminés comme ceux ayant pour coordonnées la moyenne des coordonnées des points dans le segment. Les distances entre les observations et les nouveaux centres sont déterminées et les observations sont associées, de nouveau, aux segments desquels elles sont les plus proches. Ces étapes sont répétées jusqu'à ce que les centres de segments ne bougent plus.

Le nombre de segments choisi est basé soit sur la connaissance de la base de données soit sur la méthode silhouette. De 2 à 7 segments sont choisis. Plus que 7 segments ne seraient pas distinguables visuellement. Un nombre plus grand complexifierait l'exploration visuelle. Si les connaissances sur la base de données sont limitées de façon à ce que le nombre optimal de segments ne soit pas connu, la méthode silhouette est utilisée pour déterminer le nombre de segments optimal. La méthode silhouette vise à maximiser la distance entre les segments et à diminuer la distance dans un même segment. Une fois les données regroupées en $k$ segments, l'algorithme silhouette consiste à définir une mesure silhouette $s_i$ pour chaque observation $i$ comme suit :

- $a_i$ la dissimilarité moyenne entre l'observation $i$ et tous les autres points du segment auquel appartient l'observation $i$ est calculée.
- Pour tous les autres segments $C$, $d_i^C$ la dissimilarité moyenne entre $i$ et toutes les observations appartenant à $C$ est calculée. $b_i = \min_C d_i^C$ est déterminée et peut être vu comme la distance entre $i$ et son plus proche segment.
- $s_i$ est déterminée comme suit : $s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$.

Évidemment, les observations qui sont placées seules dans un segment ont un $s_i = 0$. La dissimilarité est mesurée par la distance Euclidienne. Les observations qui ont un critère silhouette élevé qui s'approche de 1 sont très bien classées contrairement à celle qui ont un

$s_i$ proche de 0 qui peuvent être placées sur les limites de 2 segments. Pour choisir le meilleur nombre de segments selon le critère silhouette, les étapes suivies sont :

- Le nombre de segments est fixé entre 2 et 7.
- $s_i$ est déterminé pour chaque observation selon la démarche expliquée.
- $s^k$ est calculée : $s^k = \frac{\sum_i s_i}{n}$.

Le nombre de segments $k$ qui maximise $s^k$ est choisi.

Pour certaines bases, 3 était le meilleur choix et pour d'autres, les données sont reparties en 7 segments. Une fois les segments déterminés et les limites de la BOZ calculées, différentes couleurs sont associées aux segments de fonctionnement. Les zones de fonctionnements sont visualisées de 2 façons différentes. Ceci dépend de la clarté de la visualisation dépendant principalement de la base de données, ainsi, des tests peuvent être faits avant de choisir le modèle final.

- soient des polylignes avec des couleurs différentes en fonction du segment auquel appartient l'observation Les polylignes sont représentées avec un certain niveau de transparence. Grâce à la transparence, nous pouvons distinguer les zones ou plusieurs segments sont superposés. Également, la transparence permet d'associer plus d'opacité aux zones les plus denses (ou il y a plus de points).
- soient des polygones qui s'étendent sur la zone entre le minimum et le maximum de chaque point du segment. Pour ce type de visualisation, les données de chaque segment sont projetées comme expliqué dans la section 4.1. La matrice de points obtenue correspond aux points du segment $C_k$ projetée dans l'espace bidimensionnel. Les limites inférieure et supérieure du segment $C_k$ sont déterminées de la même façon que les limites de la BOZ, soient par le minimum et le maximum de chaque colonne de la matrice de points projetés. Comme pour les limites de la BOZ, les limites de chaque segment de données sont ajustées pour éviter le surajustement aux données d'apprentissage en rajoutant et en enlevant un petit pas $\epsilon_s$.

Ainsi, les cartes BOZ sont définies par les limites de la BOZ et par les limites des différents segments de fonctionnement. La classification des nouvelles observations peut se faire manuellement par l'utilisateur ou encore automatiquement comme l'explique la section 4.5. La zone de fonctionnement est définie a l'aide de certains paramètres qui sont $\epsilon_l$, la distance des limites de la BOZ, $s$ le seuil $s$ de définition des zones vides, $\epsilon_s$ la distance d'ajustement des limites des zones vides et $\epsilon_c$ la distance d'ajustement des limites de segments de fonctionnement (pour la visualisation a l'aide des polygones et pour la classification automatique).

## 4.4 Validation croisée

Tel qu'expliqué, les paramètres de définition des limites de la BOZ, des limites des segments de fonctionnement et du seuil à partir duquel une zone vide est considérée comme une zone de défauts sont réalisés à l'aide de la technique de validation croisée. Les données historiques sont réparties en données fonctionnelles et défauts. Pour faire une réplication de validation croisée, l'ensemble d'apprentissage de données fonctionnelles et de défauts est divisé en $kf$ sous ensembles. $kf - 1$ sous ensembles de données fonctionnelles sont utilisés pour trouver les limites de la BOZ, ainsi que les segments de fonctionnement. Le sous-ensemble restant de données fonctionnelles est combiné avec des données représentant des défauts pour donner le sous ensemble de test. Le nombre de données fonctionnelles versus le nombre de défauts utilisés pour le test sont égaux de façon a ce que le test donne autant d'importance aux fausses alarmes qu'à la non-détection de défauts. Pour chaque $(\epsilon_l, \epsilon_c, s, \epsilon_s)$, $r$ réplications de la démarche précédente sont répétées. La moyenne du taux de classification correcte des $r$ réplications est calculée. Le triplet $(\epsilon_l, \epsilon_c, \epsilon_t)$ qui maximise le taux de classification correcte est choisi. Une fois les paramètres $(\epsilon_l, \epsilon_c, \epsilon_t)$ déterminés, le modèle de carte de contrôle BOZ est développé tel qu'expliqué dans les sous-sections 4.2.1 et 4.3 en fonction de toutes les données fonctionnelles historiques. Les étapes d'une réplication de validation croisée sont illustrées par la figure 4.9.

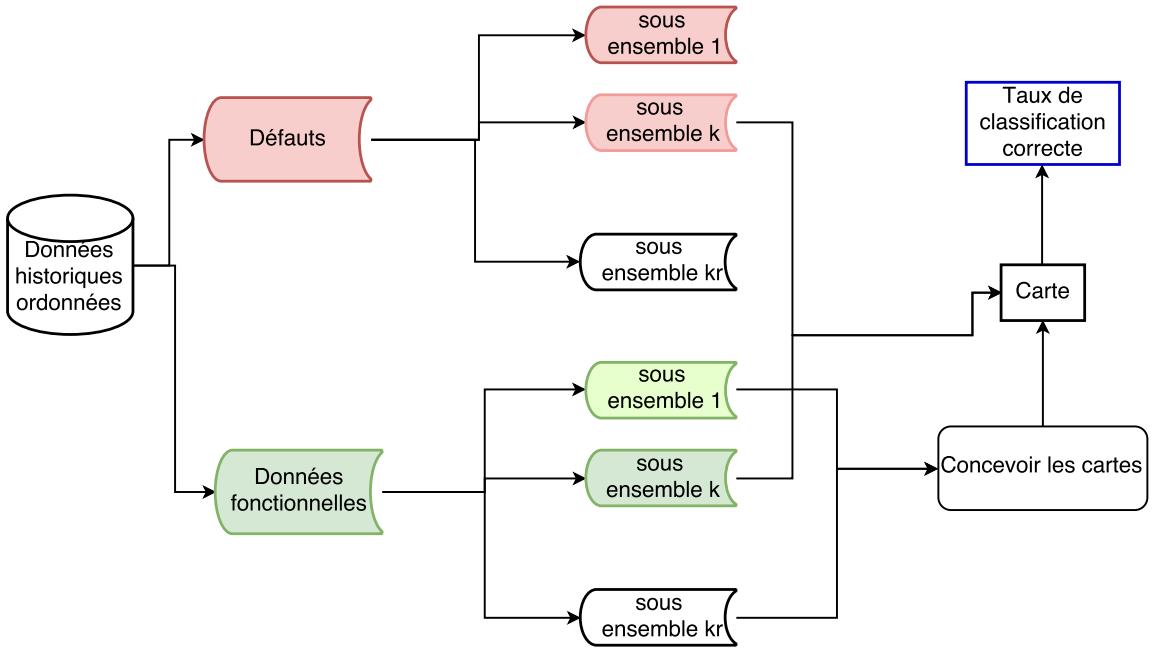


Figure 4.9 Étapes d'une réplication de validation croisée.

## 4.5 Automatisation de la détection de classification des nouvelles observations

Les cartes de contrôle ont été, au début proposées pour détecter les dérives visuellement par l'utilisateur. Or de cette façon, les cartes de contrôle peuvent être uniquement considérées comme un outil de diagnostic. En effet, si un défaut est détecté, l'expert peut utiliser les cartes de contrôle pour diagnostiquer le problème. Si, nous voulons utiliser les cartes pour la détection de défauts et vu que la détection avec les cartes sous cette forme est manuelle, la tache requiert la présence d'un humain qui utilise les cartes comme support. La détection de défauts doit être faite en continu, la vérification visuelle de chacune des observations de test peut s'avérer longue et fastidieuse. Ainsi, l'automatisation de l'algorithme de détection de défauts est essentielle. D'un autre côté, l'automatisation facilite le choix des paramètres avec l'algorithme de validation croisée.

Pour automatiser le processus de classification des nouvelles observations, l'algorithme vérifie 3 propriétés :

— si la nouvelle observation dépasse les limites de la BOZ. Pour cette étape, la nouvelle observation est standardisée $x_k^t$, projetée dans le plan Cartésien, ensuite, comparée aux limites supérieures et inférieures de la BOZ. La standardisation correspond a :

$$x_{ik}^{ts} = \frac{x_{ik}^t - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

La nouvelle observation est standardisée avec les minimums et les maximums des observations historiques, $\max\{x_i\}$ et $\min\{x_i\}$. Ensuite, le vecteur obtenu est comparé aux limites de la BOZ. S'il y a un dépassement, la nouvelle observation est considérée comme un défaut.

— si la nouvelle observation passe à l'intérieur d'une zone vide. La même nouvelle observation standardisée est comparée aux ordonnées des zones vides. Si les ordonnées du point en question sont comprises entre les coordonnées de la zone vide, alors elle est considérée comme un défaut.

— si la nouvelle observation n'appartient pas à un segment de bon fonctionnement. Ici, la nouvelle observation est comparée aux maximum et aux minimum de chacun des polygones limitant les zones de fonctionnement (segment). Si l'observation n'est à l'intérieur d'aucun des segments, au long du chemin de fonctionnement, alors, elle est considérée comme un défaut.

La classification correspond tout à fait à ce qu'un utilisateur peut apercevoir visuellement. L'automatisation de la classification des nouvelles observations est utile pour les industriels qui souhaitent implémenter l'outil de maitrise de processus développé. Ainsi, au lieu de vérifier

chacune des cartes de contrôle, l'ingénieur qualité peut se fier à la classification automatique et diagnostiquer le défaut détecté visuellement avec les cartes BOZ.

## 4.6 Évaluation de la performance de la carte de contrôle développée

Nous avons présenté dans la section 2.1.3 quelques critères de performance des cartes de contrôle. Dans cette thèse, le critère de performance choisi pour évaluer la carte de contrôle proposée est le temps opérationnel moyen ou encore ARL. Ce critère est tel que discuté largement utilisé pour évaluer les cartes de contrôle et doit être calculé sous 2 conditions possibles pour un processus de production :

— Le processus sous contrôle

— Le processus hors contrôle.

L'ARL du processus sous contrôle modélise le nombre d'observations visualisées avant d'envoyer une fausse alarme. Autrement, c'est le nombre moyen d'observations avant de détecter un défaut alors que le processus est sous contrôle. Cependant, l'ARL hors contrôle est le nombre d'observations visualisées avant de détecter un défaut lorsque le processus est hors contrôle. Ainsi, l'ARL sous contrôle doit être élevé alors que l'ARL hors contrôle doit être le plus petit possible. Dans ce travail, puisque les fonctions de distributions des observations et des limites de contrôle ne sont pas connues, le calcul théorique des $ARL_1$ et $ARL_0$ peut s'avérer complexe. Ainsi, l'ARL est estimé par des simulations de Monte Carlo.

Pour réduire le temps de calcul, les données générées pour les simulations sont des données à 3 dimensions. Les données historiques fonctionnelles $\mathbf{x}_{ic} = (x_1, x_2, x_3)$ sont générées selon une loi normale multidimensionnelle.

$$\mathbf{x}_{ic} \sim \mathcal{N}(\mu_0, \mathbf{\Sigma})$$

ou $\mu_0 = (0, 0, 0)$ et

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \sigma & \sigma \\ \sigma & 1 & \sigma \\ \sigma & \sigma & 1 \end{pmatrix} \tag{4.5}$$

Les simulations sont réalisées pour différentes valeurs de $\sigma$, pour étudier différents cas de dépendance entre les variables. Les données historiques hors contrôle $X_{hc}$ sont également générées à partir d'une loi Normale à 3 dimensions, mais avec une moyenne qui dévie de la moyenne des données sous contrôle. En se basant sur les données historiques générées, les cartes sont conçues comme expliquées dans la section 4.2. Ensuite, une base de données de test est générée pour évaluer l'ARL. Cette base est générée à l'aide d'une loi Normale

multidimensionnelle de même matrice de variance-covariance :

$$\mathbf{x}_t \sim \mathcal{N}(\mu_1, \mathbf{\Sigma})$$

ou $\mu_1 = \mu_0 + ds$. $ds$ varie pour illustrer différentes distances de dérives par rapport au processus sous contrôle. Pour chaque valeur de $ds$, 1000 bases de données sont simulées. La longueur (Run length ; RL) est mesurée. Entre autres, le numéro de l'observation correspondant au premier défaut détecté est déterminé. L'ARL est la moyenne des RL mesurés pour chaque valeur de $d$. $d$ varie de 0 à 4. Lorsque $d = 0$, nous mesurons $ARL_0$, c'est-à-dire l'ARL sous contrôle. Pour les autres valeurs de $d$, nous mesurons l'ARL hors contrôle ; $ARL_1$.
L'ARL est, alors, mesurée en fonction de 2 variables soient la distance par rapport à la moyenne sous contrôle et la dépendance entre les variables modélisée par $\sigma$. Pour conclure, la conception des cartes de contrôle BOZ passe par l'arrangement des attributs, le choix des paramètres du modèle par la méthode de validation croisée, la détermination des limites de la BOZ et la segmentation de la BOZ. Les nouvelles observations sont standardisées ensuite, classées automatiquement et visualisées pour le diagnostic.

## 4.7 Graphes de densité en coordonnées parallèles

Cette section explique la méthodologie de conception de cartes de contrôle basées sur les graphes de densité. L'objectif de cette étape est tel qu'indiqué dans le chapitre 1, de faire face aux problèmes de collecte de données rencontrés par certaines compagnies. Ces compagnies parmi lesquelles la compagnie partenaire de ce projet ont un processus long et complexe pour obtenir les données. De plus, lorsqu'elles sont collectées, les données nécessitent un prétraitement qui à son tour peut être long et complexe.

Nous avons comme but, dans cette section, d'examiner la possibilité d'utiliser les graphes de densité en coordonnées parallèles afin d'aider et de soutenir la maitrise de processus de production par l'exploration visuelle des données industrielles. La démarche de développement des cartes de contrôle densité est décrite par la figure 4.10.

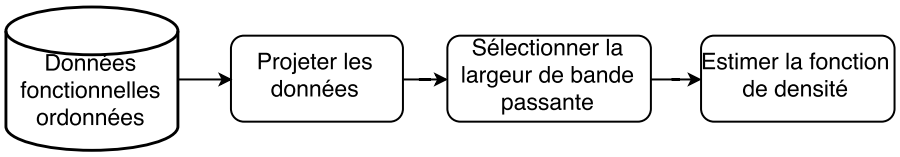


Figure 4.10 Étapes de conception des cartes BOZ.

Le développement des graphes de densité en coordonnées parallèles nécessite une base de données historique incluant des observations dans une classe d'intérêt, i.e. des données sous contrôle. Cette démarche s'applique après l'arrangement des variables présenté dans le chapitre 3. L'utilité du réarrangement est de même que dans la procédure de conception des cartes BOZ. Les données sont ordonnées dans l'objectif de visualiser les dépendances. La séparation des données n'est pas un critère d'intérêt étant donné que nous ne nous intéressons pas à la segmentation des données dans cette version. L'arrangement des variables vise à afficher au mieux les relations entre les variables et à mettre en valeur le pattern général suivi par les données fonctionnelles. La conception des cartes de contrôle densité est basée sur les données fonctionnelles ordonnées et se fait en 2 grandes étapes :

— La projection des données standardisées en coordonnées parallèles dans un plan Cartésien bidimensionnel.

— L'estimation de la fonction de densité des données obtenues par la méthode du noyau.

En effet, le graphe en coordonnées parallèles est transformé en graphes a $2D$ de façon à ce que plusieurs points appartenant aux polylignes sont représentées à la place des polylignes

elles mêmes. La matrice de projection appliquée à toutes les paires de variables adjacentes donne la matrice de projection de tous les points en coordonnées parallèles, soit :

$$\mathbf{Y}_{tot} = \begin{pmatrix} f_1^1(\frac{1}{n_e}) & \cdots & f_1^1(1) & f_1^2(\frac{1}{n_e}) & \cdots & f_1^2(1) & f_1^{(p-1)}(\frac{1}{n_e}) & \cdots & f_1^{(p-1)}(1) \\ f_2^1(\frac{1}{n_e}) & \cdots & f_2^1(1) & f_2^2(\frac{1}{n_e}) & \cdots & f_2^2(1) & f_2^{(p-1)}(\frac{1}{n_e}) & \cdots & f_2^{(p-1)}(1) \\ \vdots & \vdots & \ddots & & \vdots & \vdots & & & \\ f_n^1(\frac{1}{n_e}) & \cdots & f_n^1(1) & f_n^2(\frac{1}{n_e}) & \cdots & f_n^2(1) & f_n^{(p-1)}(\frac{1}{n_e}) & \cdots & f_n^{(p-1)}(1) \end{pmatrix} \tag{4.6}$$

La matrice $\mathbf{Y}_{tot}$ est utilisée pour estimer la fonction de densité. Avant de passer à l'estimation de la fonction de densité, la matrice $\mathbf{Y}_{tot}$ est transformée en un vecteur, comme elle représente une seule variable $a_2$ du plan Cartésien. Ainsi, nous notons

$$\mathbf{y}_2 = \Big[f_i^k(\frac{j}{n_e})\Big].$$

Les abscisses des points représentées dans la figure 4.11 sont regroupées dans un vecteur $\mathbf{y}_1$ entre 1 et $d$ avec un pas $\frac{1}{n_e}$, ainsi,

$$\mathbf{y}_1 = \Big[k + \frac{j}{n_e}\Big]$$

ou $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, p-1\}$ et $j \in \{1, \ldots, n_e\}$.
Pour toute la suite et pour l'estimation de la fonction de densité, nous utilisons la matrice composée des vecteurs $\mathbf{y}_1$ et $\mathbf{y}_2$ que nous notons $\mathbf{X} = (\mathbf{y}_1, \mathbf{y}_2)$. La fonction de densité des coordonnées des points du graphe de droite de la figure 4.11 est estimée à l'aide de la méthode d'estimation de densité par noyau. La méthode par noyau est une méthode non paramétrique largement utilisée.
Soit $\mathbf{x}_j$ un vecteur représentant une observation multidimensionnelle de $\mathbf{X}$. Pour le cas des données multivariées, la fonction de densité du noyau est estimée comme suit :

$$\hat{f}_{\mathbf{H}}(\mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}_j - \mathbf{x}_i)$$

où $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jd})^T$ et $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ pour $i \in \{1, \ldots, n\}$ sont des vecteurs d'observations. Dans ce cas, $d$ qui est la dimension du vecteur $\mathbf{x}_j$ est égale à 2.

$$K_H(\mathbf{x}_j) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x}_j)$$

$\mathbf{H}$ est la bande passante pour $\mathbf{x}_j$ et $K$ est la fonction du noyau. La fonction noyau est une fonction symétrique autour de 0, positive, paire et bornée. La fonction noyau satisfait

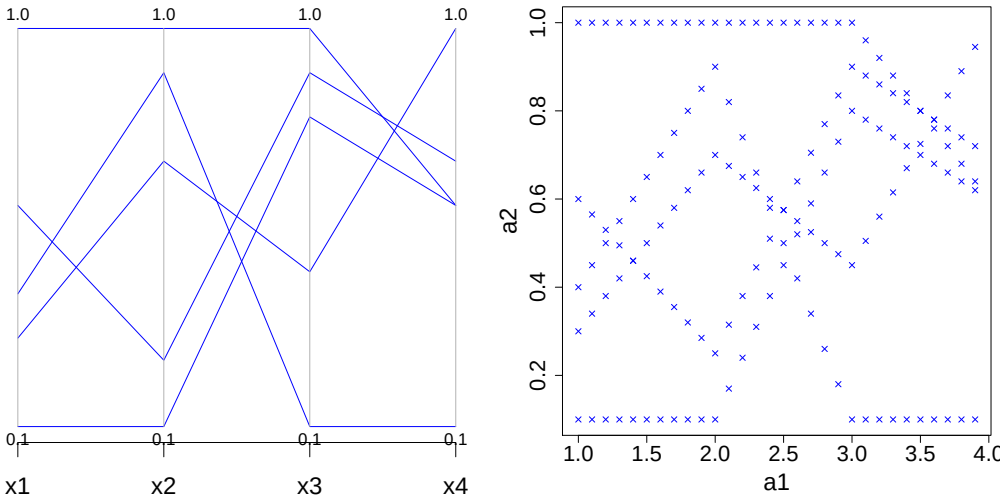


Figure 4.11 Données multidimensionnelles projetées dans un plan Cartésien bidimensionnel.

également aux conditions :

— $K_H$ est une fonction de densité $\int_{-\infty}^{+\infty} K(\mathbf{x})d\mathbf{x} = 1$

— Le noyau est de carré intégrable $\int_{-\infty}^{+\infty} K^2 d\mathbf{x} \leq \infty$

— Le noyau admet un moment d'ordre $\int_{-\infty}^{+\infty} \mathbf{x}^2 K(\mathbf{x})d\mathbf{x} = 0$

Parmi les fonctions noyaux les plus utilisées, nous trouvons le noyau uniforme qui correspond à la fonction de densité de la loi uniforme, le noyau triangulaire, et le noyau Gaussien qui correspond à la fonction de densité de la loi Normale :

$$K(\mathbf{x}) = \frac{1}{(2\pi)|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}.$$

Nous suggérons pour cette thèse l'utilisation d'un noyau Gaussien. À partir d'une vérification visuelle, nous vérifions le choix de ce noyau et nous le changeons au besoin en une loi uniforme ou triangulaire par exemple. Cependant, le choix d'un noyau Gaussien est celui qui peut s'adapter mieux à différentes distributions de données. La vérification visuelle se fait en comparant le graphes en coordonnées parallèles au graphe de densité. Un aspect important de l'estimation de densité par noyau est le choix de $\mathbf{H}$. $H$ est un paramètre de lissage, il est appelé la largeur de la bande passante. Dans le cas de données multidimensionnelles, $\mathbf{H}$ est une matrice. Le choix des valeurs de la bande passante affecte la forme générale de la fonction de distribution. En effet, si les valeurs de $\mathbf{H}$ sont trop petites, alors, l'estimation devient trop sensible au bruit. Cependant, si les valeurs de $\mathbf{H}$ sont trop grandes, la fonction estimée perd les traits essentiels. Il existe également plusieurs méthodes pour estimer la matrice $\mathbf{H}$. Un choix simple de $\mathbf{H}$ est une matrice diagonale avec les mêmes valeurs de la diagonale, soit $h^2$ ou $h$ est un paramètre de lissage. Silverman (1986) suggère une règle de pouce pour fixer la

bande passante univariée de chacune des variables $x_i$ :

$$h_i = 1.06\hat{\sigma}_i$$

ou $\sigma_i$ est l'écart type de la variable $x_i$. À partir de ce paramètre uni varié, nous définissons la matrice de bande passante $\mathbf{H}$ telle que :

$$\hat{\mathbf{H}} = \begin{pmatrix} \hat{h_1} & 0 \\ 0 & \hat{h_2} \end{pmatrix} \tag{4.7}$$

La matrice de bande passante est de dimensions $(2, 2)$ car nous utilisons la matrice $\mathbf{X}$ des points projetés dans le plan Cartésien. La règle de Silverman fonctionne bien lorsque la fonction noyau est une fonction Gaussienne. Ainsi, lorsque le noyau Gaussien est utilisé, nous utilisons la règle de Silverman pour estimer le paramètre de lissage. Cependant, lorsque nous changeons le noyau, cette estimation n'est plus trop fiable. Une deuxième règle est utilisée qui est la règle de Scott qui définit les éléments diagonaux de la matrice $\mathbf{H}$ comme suit :

$$\hat{h_i} = n^{-1/(d+4)\hat{\sigma}_i} = n^{-1/6\hat{\sigma}_i}.$$

Ce qui peut être remarqué est que les paramètres de lissage sont souvent définis en fonction de l'estimée de l'écart type. Dans ce travail l'objectif est de proposer des cartes de densité pour diagnostiquer et détecter les défauts. L'estimation se fait en fonction des vecteurs résultant de la projection. Ainsi, si $\mathbf{y}_2$ de la matrice $\mathbf{X}$ peut prendre différentes valeurs entre 0 et 1, les valeurs de l'abscisse $\mathbf{y}_1$ sont prédéfinies par le pas choisi ou encore le nombre d'échantillons $n_e$ entre 2 axes parallèles (discuté dans la section 4.1). Ainsi, le meilleur noyau choisi pour estimer la distribution univariée de $\mathbf{y}_1$ est un noyau de loi uniforme. Cependant, le noyau est bivarié, alors, nous gardons le même choix de noyau Gaussien. Mais, nous tenons compte de la distribution de $\mathbf{y}_1$ dans le choix de la bande passante correspondant à la variable $\mathbf{y}_2$.
$H$ est, alors, définie comme suit :

$$\hat{\mathbf{H}} = \begin{pmatrix} \hat{h_1} & 0 \\ 0 & \hat{h_2} \end{pmatrix} \tag{4.8}$$

ou $h_1$ est estimée par la règle de Scott et $h_2$ est estimée par la règle de Silverman. Pour plus de détails, Duong (2007) est une source d'intérêt où l'estimation par noyau fourni par le logiciel $R$ est expliquée. Ce logiciel $R$ est utilisé pour implémenter l'algorithme expliqué. Ayant comme objectif de proposer une carte qui s'adapte avec un nombre assez limité de données, le temps requis pour générer les cartes n'est pas long, ainsi, nous supposons que l'ajustement de ces paramètres après une vérification visuelle reste possible. Une fois, la fonction de densité est

estimée, elle est représentée en utilisant différentes couleurs selon les niveaux de densité. Les couleurs vont du rouge au bleu. La couleur rouge visualise la zone non dense tandis que la couleur bleue visualise la zone dense. Lors de l'acquisition de nouvelles observations, celles-ci sont standardisées, comme suit :

$$x_{ik}^{ts} = \frac{x_{ik}^{t} - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

ou $\max\{x_i\}$ et $\min\{x_i\}$ sont le maximum et le minimum de la variable $x_i$ et $k \in \{1, \ldots, p\}$. Ensuite, les nouvelles observations standardisées sont représentées par dessus le graphe de densité en coordonnées parallèle. Les zones les plus denses représentent l'équivalent de la BOZ de la carte BOZ. Pour classer les nouvelles observations, il suffit de regarder les zones par lesquelles elles passent. Si elles passent par des zones rouges, alors, elles sont considérées comme défauts et si elles passent uniquement par des zones bleues ou jaunes, alors elles sont classées fonctionnelles.

## 4.8 Conclusion

Dans cette section, nous présentons 2 de cartes de contrôle qui s'adaptent avec la disponibilité des données historiques. Les 2 cartes sont basées sur des données ordonnées avec l'algorithme présenté dans le chapitre 3. La conception des cartes se base sur la caractérisation de la zone de fonctionnement avec des techniques de projection géométrique combinées avec des techniques d'apprentissage statistique. Les cartes sont évaluées avec différentes bases de données (cf. chapitre 5).

# CHAPITRE 5 RÉSULTATS

Dans le chapitre 5, nous évaluons les différents algorithmes proposés dans le chapitre 3 et le chapitre 4. D'abord, l'algorithme de réarrangement des attributs est évalué dans la section 5.1. Dans un deuxième temps, les cartes BOZ sont évaluées avec une base de données simulées et une base de données réelles et comparée aux cartes d'Hotelling. Également, l'impact de l'algorithme d'arrangement des variables sur les résultats de détection de défauts par les cartes BOZ est évalué. Ensuite, les cartes de contrôle densité sont présentées et discutées. Ces cartes sont également comparées aux cartes de BOZ et aux réseaux de neurones.

Tout au long de la section résultat, différentes bases de données sont utilisées. L'objectif de variation de bases de données est d'illustrer les différents concepts présentés dans la thèse. Les bases de données sont choisies en fonction de l'algorithme.

## 5.1 Évaluation de l'ordre des variables en coordonnées parallèles

Dans le chapitre 3, nous avons présenté un cadre général de réarrangement de variables selon un objectif prédéfini. Cet objectif détermine le critère de réarrangement. Pour évaluer différentes métriques de réarrangement, différentes bases de données sont utilisées. Le réarrangement est réalisé dans deux objectifs soient la dépendance entre les variables et la séparation des données. La base de données de qualité de vins blancs est utilisée, pour la métrique de dépendance et la base de données génétiques (Golub data) est utilisée pour la séparation. Pour fin de rappel, l'expression de l'information générale proposée entre 2 attributs $x_i$ et $x_j$ est donnée par l'équation suivante :

$$GI(x_i, x_j) = \frac{1}{G''(1)} \int \int G(\frac{f(x_i, x_j)}{h(xi1, x_j)}) h(x_i, x_j) \tag{5.1}$$

### 5.1.1 Critère de dépendance : Base de données de la qualité de vin

La base de données vins blancs est choisie, car elle est l'une des bases de données les plus utilisées dans l'évaluation des algorithmes de réarrangement des attributs en coordonnées parallèles.

— ***Description de la base de données des vins blancs***

Cette base de données est le résultat d'une analyse chimique de vins. Elle a été créée et offerte par Cortez *et al.* (2009) en 2009. La base créée contient des informations à propos

de vins blancs et rouges. Dans cette section, les données de vins blancs sont exclusivement utilisées. Cette base de données est utilisée, dans la littérature, pour des taches d'étude de corrélation ou encore pour l'évaluation des algorithmes de classification. Elle est, particulièrement, très populaire dans l'évaluation des algorithmes de réarrangement des attributs en coordonnées parallèles. Elle est utilisée par Dasgupta et Kosara (2010), pour tester des algorithmes d'arrangement de variables basées sur plusieurs métriques, dont le parallélisme et les angles d'intersection des polylignes. Il s'agit d'une base de données gratuite disponible sur internet [1]. Cette base de données inclut 4898 observations. Elle comporte 12 attributs qui sont l'acidité fixée ($x_1$), l'acidité volatile ($x_2$), l'acidité citrique ($x_3$), les résidus de sucre ($x_4$), les chlorures ($x_5$), le dioxyde de soufre libre ($x_6$), le dioxyde de soufre total ($x_7$), la densité ($x_8$), le pH ($x_9$), les sulfates ($x_{10}$), l'alcool ($x_{11}$) et la qualité ($x_{12}$) qui représente un score entre 0 et 10 pour illustrer la qualité du vin blanc. Nous associons des noms plus courts aux variables ($x_i$) dans le but d'éviter l'encombrement des graphiques.

— ***Réarrangement des variables pour objectif de dépendance***

Pour évaluer l'arrangement dans l'objectif de souligner les dépendances entre les attributs, $F$ et $H$ de l'équation 5.1 sont choisies telles que $F(x_i, x_j)$ est la probabilité jointe de $x_i$ et $x_j$ et $H(x_i, x_j)$ est le produit des probabilités marginales.

$$F(x_i, x_j) = p(x_i, x_j)$$

$$H(x_i, x_j) = p(x_i)p(x_j)$$

ou $(i, j) \in \{1, \ldots, n\}^2$. $F$ mesure la dépendance et $H$ mesure l'indépendance. L'intégrale de l'équation 5.1 est estimée à l'aide de la somme. L'équation équivalente est alors :

$$GI(x_i, x_j) = \frac{1}{G''(1)} \sum_i \sum_j G\left\{\frac{p(x_i, x_j)}{p(x_i)p(x_j)}\right\} p(x_i)p(x_j)$$

Pour changer la statistique tout en étudiant la dépendance, nous varions $G(.)$ en gardant les mêmes $F$ et $H$. Nous étudions l'impact de changement de la statistique sur l'ordre généré. Premièrement, l'ordre optimal obtenu à l'aide de CPLEX est comparé à l'ordre sous-optimal obtenu avec l'approche Glouton. Nous commençons par mesurer la distance entre la solution optimale proposée par CPLEX et la solution trouvée par l'algorithme Glouton par rapport à la solution optimale. La figure 5.1 illustre les 2 résultats : Dans les 2 figures, les données sont ordonnées pour objectif de dépendance à l'aide de la même statistique d'information mutuelle. La solution trouvée par donne une somme d'information générale $\sum_i \sum_j GI(x_i, x_j) = 2.53$

1. http ://archive.ics.uci.edu/ml/datasets/Wine

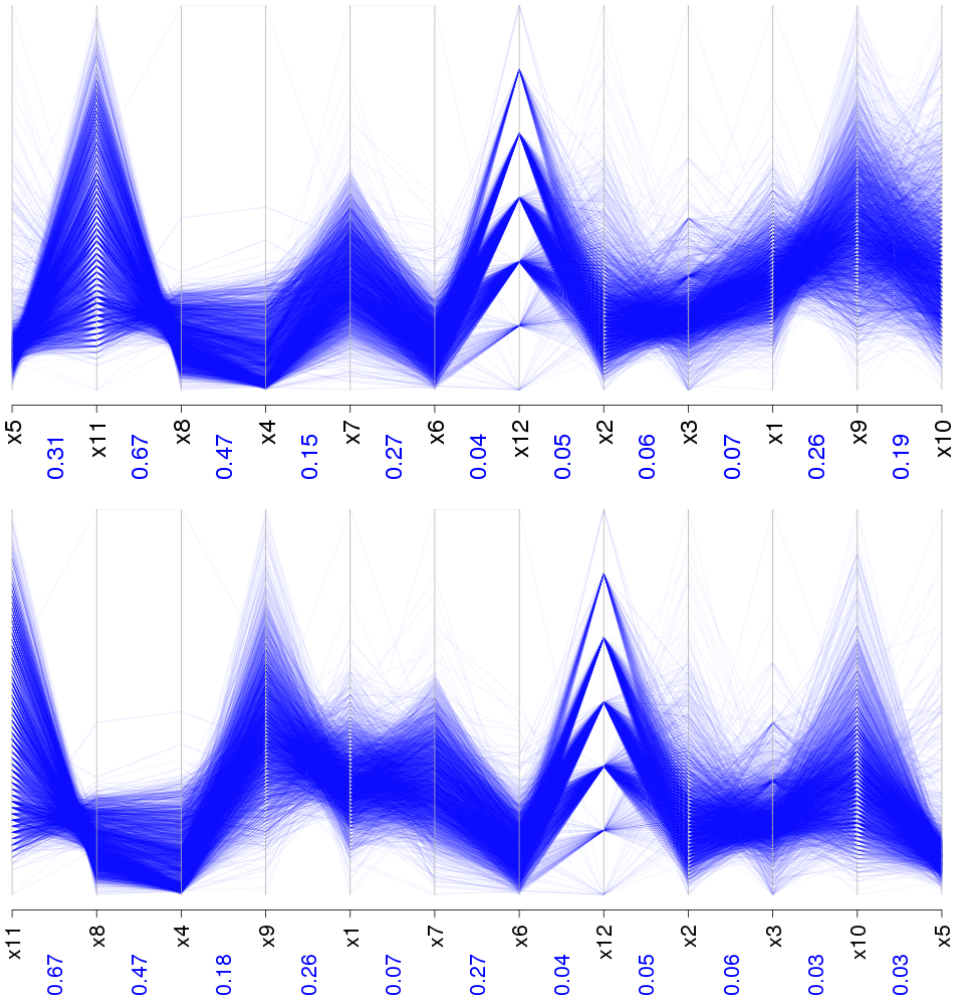


Figure 5.1 Figure montrant les données de vins blancs réordonnées avec la statistique d'information mutuelle. Les valeurs en bleu représentent les valeurs de l'information générale entre 2 paires d'attributs.

comparée à la somme de $\sum_i \sum_j GI(x_i, x_j) = 2.13$ obtenu avec l'algorithme Glouton. La différence d'information générale entre les 2 solutions est de 0.40. La différence est acceptable en tenant compte des avantages de l'algorithme Glouton. Cplex ne donne pas une liste mais une matrice d'adjacence avec 2 voisins par attributs. Le temps d'exécution de l'algorithme Glouton est légèrement inférieur que l'exécution de l'algorithme Cplex (7 secondes par rapport à 30 secondes). L'algorithme Glouton est implémenté dans $R$, alors que Cplex nécessite une connexion avec $R$. La matrice d'information générale sert comme entrée pour l'algorithme d'optimisation. Les ordres obtenus par CPLEX et par l'algorithme Glouton ne sont pas identiques. Cependant, plusieurs attributs sont placés voisins par les 2 algorithmes, par exemple, $(x_8, x_4)$, $(x_2, x_3)$, $(x_2, x_{12})$ et $(x_7, x_6)$. Nous pouvons remarquer la présence de certaines valeurs aberrantes. Essayant de supprimer ces valeurs, plus que 20% des observations devaient

être supprimées. Elles représentent les vins de qualité très élevé ou les vins de très mauvaises qualité. Ainsi, nous gardons toutes les données pour cette étude. Pour les autres statistiques, l'ordre est optimisé avec l'algorithme Glouton uniquement. Les statistiques évaluées sont Cressie-Read, Tukey-Freeman, Pearson-Khi2 et Neyman. La figure 5.2 montre les données de vins blancs réordonnées selon les différentes statistiques citées.

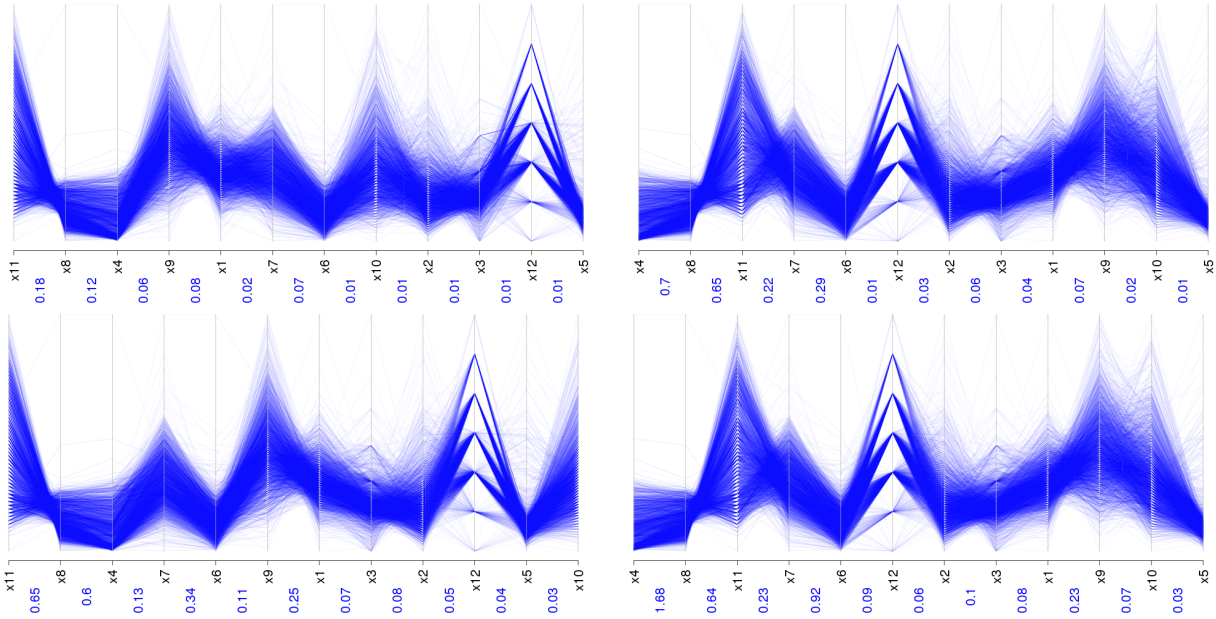


Figure 5.2 Figure montrant les données de vins blancs réordonnées avec la statistique d'information mutuelle. Les valeurs en bleu entre chaque couple d'attributs adjacents sont les valeurs de l'information générale $GI(x_i, x_j)$. De haut en bas et de droite à gauche sont représentées les données ordonnées avec la statistique Freeman-Tukey, Neyman, Cressie et Pearson.

Les statistiques Neyman et Pearson donnent exactement le même ordre d'attributs. Ce qui est attendu vu que les statistiques de Neyman et Pearson sont connexes et équivalentes. L'information mutuelle, la statistique de Cressie et celle de Tukey-Freeman propose les mêmes 3 premières variables $x_{11}$, $x_8$ et $x_4$. L'ordre continue de la même façon pour l'information mutuelle et la statistique de Tukey jusqu'à la $7^{ème}$ variable. De plus, même pour les statistiques qui ne commencent pas de la même façon, il y a toujours des variables qui sont placées dans le même voisinage par toutes les statistiques. Par exemple, $x_4$ est toujours placée au voisinage de $x_8$. Également, $x_6$ est toujours placée a coté de $x_7$. Cependant, avec un critère mis à l'échelle, nous pouvons remarquer que $\sum_i \sum_j GI(x_i, x_j)$ est plus élevée pour la statistique de Neyman et Pearson (4.18) qui donnent le même ordre, suivie par des sommes comparables pour l'information mutuelle et Cressie avec une somme autour de 2 et finalement, la statistique de Tukey-Freeman qui donne une somme de 0.58. La comparaison de $\sum_i \sum_j GI(x_i, x_j)$ peut, probablement donner une idée sur à quel point une statistique peut donner de l'infor-

mation sur la dépendance entre 2 attributs. À partir des valeurs de chacune des statistiques, nous pouvons remarquer que le rapport entre 2 attributs dépendants et 2 attributs indépendants est plus important dans le cas de Pearson et Neyman. Le rapport de la statistique de Tukey-Freeman entre les attributs les plus dépendants et les attributs les moins dépendants est de 1, tandis que ce même rapport est de 1 pour la statistique de Pearson. Donc, dans ce cas, ce cas, Pearson et Neyman distinguent mieux les variables dépendantes des variables peu dépendantes. Ceci peut probablement être expliqué par les courbes des fonctions qui permettent d'obtenir chacune des statistiques. La figure 5.3 représente ces graphes de fonctions. Nous pouvons voir que la courbe de la fonction correspondant à la statistique de Neyman

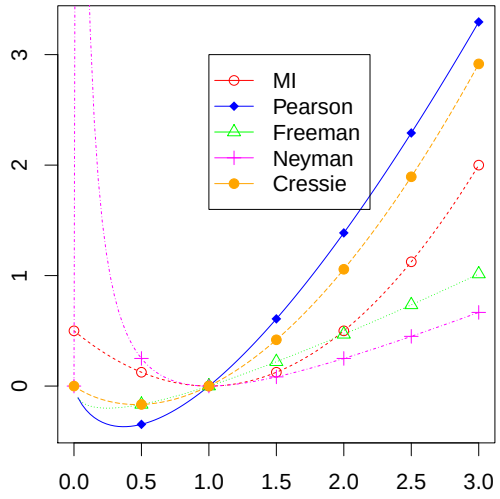


Figure 5.3 Figure illustrant le comportement des courbes des fonctions qui permettent d'obtenir les différentes statistiques étudiées autour de 1.

décroit très rapidement avant la valeur 1, ce qui permet d'associer des valeurs bien distinctes si les rapports $\frac{F}{H} < 1$ ou si $\frac{F}{H} = 1 \pm \epsilon$. La même chose se produit lorsque $\frac{F}{H} > 1$ pour la courbe de Pearson. Ce phénomène est moins présent pour les courbes de l'information mutuelle et pour la courbe de Cressie. Finalement pour la courbe de Freeman-Tukey, la courbe a une pente très faible. Ainsi, l'image de $\frac{F}{H}$ n'est pas très différente de 0 qui est l'image de $\frac{F}{H}$ lorsque $\frac{F}{H} = 1$. Nous pouvons, alors conclure que lorsque nous avons comme objectif de distinguer les variables dépendantes des variables indépendantes, nous suggérons utiliser Pearson ou Neyman. Ceci revient à ne considérer que les attributs très dépendants puisque les valeurs de l'information générale pour les attributs peu dépendants se rapprochent des valeurs de l'information générale des variables indépendantes qui est de 0.

En conclusion, le changement de la statistique donne un ordre différent même avec des fonctions $F$ et $H$ restent fixées. Cependant, en général, plusieurs relations entre les variables dépendantes sont détectées par les différentes statistiques. La section 5.1.2 présente la deuxième

application de l'algorithme d'arrangement de variables soit l'arrangement pour l'amélioration de la séparation visuelle des données.

### 5.1.2 Critère de séparation : Base de données génétique

Dans cette sous-section, nous étudions le réarrangement des variables dans l'objectif de séparer les données. L'objectif de ce type de réarrangement est de souligner la détection des segments.

— ***Description de la base de données génétiques.***

Les données génétiques appelées aussi Golub data ont été créées par Golub *et al.* (1999). Elles comportent 47 patients atteints de leucémie lymphoblastique aiguë et 25 patients atteints de leucémie myéloïde aiguë. Les observations ont été analysées avec Affymetrix Hgu6800 chips, ce qui a entraîné 7129 expressions de gènes (sondes Affymetrix). Les données ont été prétraitées, donnant 2030 attributs . Ces données sont de très haute dimension. Donc, la sélection du sous ensemble d'attributs les plus informatifs est cruciale. Dans ce cas-ci, trouver les gènes qui séparent le mieux les données est plus intéressant que trouver les gènes les plus dépendants. Pour les données Golub, nous ordonnons les attributs en fonction de la mesure de séparation discutée dans le chapitre 3.4.

— ***Organisation des attributs de la base de données génétique selon la mesure de séparation.***

Comme le nombre total de gènes est de 2030, il est impossible de les représenter tous sur un même graphe en coordonnées parallèles, vu les limites de la taille de l'écran. Nous choisissons les $q = 50$ gènes qui séparent le mieux les données selon le critère de séparation bidimensionnelle. Pour ordonner et sélectionner les données qui séparent au mieux les données, $F$ est choisi comme la fonction de répartition d'un mélange de Gaussiennes et $H$ comme la fonction de distribution d'une loi Gaussienne unique tel qu'expliqué dans le chapitre 3. Étant donné que l'ajustement d'un mélange de Gaussiennes est long, les paramètres du mélange de Gaussiennes sont estimés à l'aide de la méthode de segmentation des $k$ moyennes. Le nombre de segments optimal n'est pas connu, ainsi, nous choisissons comme suggéré dans la méthodologie un nombre de segments élevé soit $k = 7$. Le choix de $G(.)$ n'est pas aussi important que le choix de $F$ et $H$, alors, $G(u) = ulog(u)$ est utilisée. Pour tester la possibilité de résolution de ce type de problème avec CPLEX, nous avons essayé de réordonner les données avec comme critère la corrélation de Pearson. La matrice de corrélation de Pearson est utilisée pour tester la possibilité de résoudre un tel problème avec le solveur CPLEX. Elle est utilisée, car elle ne nécessite pas un temps de compilation long comme la matrice d'information générale de la mesure de séparation. La mesure de séparation est longue à calculer même avec l'algorithme d'approximation basé sur la méthode des $k$ moyennes. Même pour l'algorithme Glouton, nous

proposons une méthode alternative détaillée dans la suite. Imposant la sélection de $q = 50$ attributs, CPLEX n'a pas convergé. Donc, il est impossible de trouver la solution optimale avec CPLEX pour ce type de problèmes. Ainsi, l'algorithme Glouton est utilisé pour trouver l'ordre optimal séparant les données. Le calcul de la matrice d'information générale est très lent vu qu'il faut calculer un nombre égal à $C_2^{2030}$ de mesure de séparation même si la résolution se fait avec l'algorithme Glouton. Pour éviter ce calcul, la première variable de la liste est sélectionnée comme étant celle qui maximise la séparation des données selon un critère univarié. Les autres variables sont choisies avec l'algorithme Glouton.
En visualisant les données réparties en 7 segments, nous avons remarqué que 3 parmi eux peuvent être regroupés. Ceci évite le calcul long de la matrice d'information générale et ne doit pas affecter autant le résultat final. En conséquence, 4 segments sont visualisés. La qualité de détection des segments est évaluée et comparée avec la qualité dans le graphique ou les données sont réordonnées avec la corrélation de Pearson. Les résultats de réarrangement sont présentés sur la figure 5.4. Les données sont segmentées et chaque segment est visualisé par une couleur différente pour illustrer la séparation.

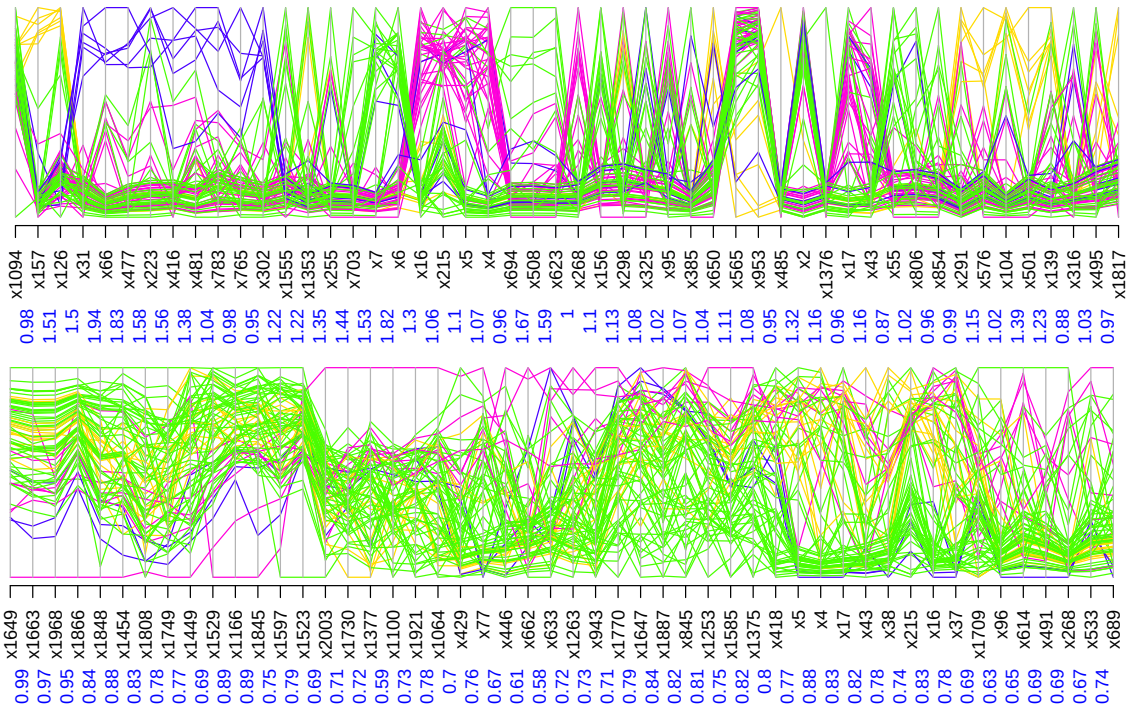


Figure 5.4 Figure montrant les données génétiques réordonnées avec le critère de séparation (figure en haut) et avec la corrélation de Pearson (figure d'en bas).

À partir de la figure 5.4, les segments ne sont pas distinguables lorsque les données sont réordonnées avec la corrélation de Pearson. Cependant, nous pouvons distinguer 4 segments

lorsque le critère de réarrangement est la séparation, les segments en bleu, magenta, jaune et vert. Ces segments sont séparables entre certaines variables et mélangés avec les autres segments entre certaines autres variables. Par exemple, de $x126$ jusqu'à $x1555$, le segment en bleu est le seul distinguable parmi les autres. Entre les variables $x1094$ et $x126$, le segment en jaune est distinguable. Et entre les variables $x6$ et $x694$, le segment en magenta est séparable des autres. Ainsi, nous pouvons caractériser chaque segment au niveau de certaines variables.

Ceci est confirmé avec l'information totale de la mesure de séparation évaluée pour les 2 ordres présentés. En effet, l'ordre obtenu avec la corrélation de Pearson donne une somme $\sum_i \sum_j GI(x_i, x_j) = 57$ et l'ordre obtenu avec le critère de séparation donne une somme $\sum_i \sum_j GI(x_i, x_j) = 30$. Cet ordre est toutefois moins bon que celui qui aurait était proposée par l'algorithme Glouton de départ. En effet, la valeur du critère de la première paire de données n'est pas la plus élevée. Contrairement à l'algorithme Glouton de départ, l'algorithme modifié ne place pas les 2 premières variables comme celles qui maximisent le critère bidimensionnelle de séparation. Toutefois, cette mesure est comparable aux mesures des autres paires sélectionnées. Avec cette contrainte, l'ordre proposé permet, quand même, de détecter les segments de données visuellement.

Ainsi, le critère de séparation améliore la détection des segments et la séparation visuelle des données. Les deux sections précédentes présentent une évaluation de l'algorithme d'arrangement des variables avec 2 bases de données différentes. L'évaluation montre alors que l'algorithme permet une meilleure exploration visuelle des données selon le propos d'arrangement. Dans la première application, la base de données contient 12 variables qui sont toutes ordonnées. Dans ce cas, l'utilisation de CPLEX est possible. Cependant, dans la deuxième application, 50 variables sont ordonnées parmi 2030. Dans ce cas, l'algorithme d'optimisation avec CPLEX ne converge pas. Les 2 bases de données présentées ne sont pas en lien avec le contrôle qualité. Elles ne sont pas utilisées dans l'évaluation des cartes de contrôle proposées. Ainsi, 2 autres bases, en lien avec les processus industriels de production sont utilisées pour l'évaluation. Les ordres donnés par chacun des critères discutés sont présentés. Ceci a pour objectif d'évaluer l'impact du réarrangement des variables dans le développement des cartes de contrôle.

## 5.2 Évaluation des cartes de contrôle basées sur la BOZ

Dans cette section, nous évaluons l'outil par un cas simple où les données sont simulées, une étude de cas qui sert à étudier et analyser les concepts clefs de l'approche proposée, ensuite par un cas de données réelles pour évaluer la performance de l'outil. Bien que l'ordre des variables est étudié dans la section 5.1, dans la section suivante, nous présentons pour chacune des bases de données l'ordre proposé, discutons l'importance de l'ordre des variables dans la détection des défauts et dans la réduction du taux de fausses alarmes.

### 5.2.1 Présentation et explication des cartes de contrôle : base de données simulées

Afin d'expliquer les cartes de contrôle proposé, une base de données simulées est utilisée. La description commence par une explication des limites de la BOZ, des types de défauts détectées avec ces limites de la BOZ et ensuite avec les segments de fonctionnement. Une comparaison avec la carte $T^2$ d'Hotelling est présentée. L'utilisabilité de la carte est étudiée à travers une expérience d'utilisateurs.

— ***Description de la base de données simulées***

Une base de données simulant un processus de formage-remplissage-scellage est générée. Les attributs considérés sont 3 températures ponctuelles ($T1$, $T2$ et $T3$), un temps de refroidissement ($T.Cool$), un temps de remplissage ($Dep.T$), un débit de pression ($D.Press$), un débit d'eau ($D.water$), un débit de gaz froid ($D.Cool$) et finalement un indicateur de niveau ($I.Vol$). Les données sont générées de telle façon que certaines variables sont dépendantes et d'autres sont indépendantes. Premièrement, les 3 attributs $I.Vol$, $D.Cool$ et $T.Cool$, sont aléatoirement et indépendamment générés respectivement entre 197 et 202, 0.2 et 0.94 et 26 et 83. En fonction de $I.Vol$ et $D.Cool$, 6 autres variables sont générées comme l'indique le tableau 5.1. Comme $D.Press$, $T1$ et $D.water$ sont générées en fonction de $I.Vol$, elles sont toutes interdépendantes. Ceci est valide pour $T2$, $T3$ et $Dep.T$ qui sont toutes générées en fonction de $D.Cool$. Les relations entre les différents attributs sont résumées dans le tableau 5.1. Différents types de dépendances sont utilisées pour générer les attributs, par exemple, des relations linéaires, elliptiques, puissance, quotient ou encore racine carrée. Un bruit blanc a été rajouté à chacun des attributs dans le but de s'approcher du cas réel industriel.

Des défauts sont également générés pour tester les cartes proposées. 2 types de défauts sont générés :

— Des données qui ne respectent pas les limites supérieures ou inférieures, une variable aléatoire est rajoutée au maximum de certaines variables ou un pas est soustrait du

Tableau 5.1 Relations qui relient les différents attributs. Par exemple la ligne 4, colonne 6 montre la relation qui relie $D.Cool$ et $T2 : T2 = 18D.Cool^2 + 34$.

| | $D.Press$ | $T1$ | $D.water$ | $T2$ | $T3$ | $Dep.T$ |
|---|---|---|---|---|---|---|
| $I.Vol$ | $0.3I.Vol+0.15$ | $-I.Vol + 250$ | $\frac{1}{I.Vol+4}$ | | | |
| $D.Press$ | | $-3.3D.Press+250.5$ | $\frac{0.3}{D.Press+1.05}$ | | | |
| $T1$ | | | $\frac{10}{90-T1}$ | | | |
| $D.Cool$ | | | | $18D.Cool^2 + 34$ | $49D.Cool^4+10$ | $-9\sqrt{D.Cool} + 10$ |
| $T2$ | | | | | $0.15(T2 - 34)^2 + 10$ | $-4.36(T2 - 34)^{1/4} + 10$ |
| $T3$ | | | | | | $-5.53(T3 - 10)^{1/8} + 10$ |

minimum d'une variable.

— 2 attributs supposés indépendants sont générés aléatoirement, cette simulation permet de générer des défauts qui ne suivent pas la structure des données, ou ne respectent pas les relations entre les variables, c'est à dire, les segments de données ou les limites de relations entre les variables.

1000 observations historiques sont générées dont 800 sont fonctionnelles et 200 représentent des défauts. Ces données sont utilisées pour concevoir les cartes BOZ, fixer les paramètres du modèle, etc. 440 observations sont également générées pour tester les cartes. La moitié des données de test est fonctionnelle et l'autre moitié représente des défauts. Nous commençons par discuter l'ordre trouvé par chaque critère soit séparation et dépendance, pour continuer l'évaluation de l'outil. Nous terminons par une évaluation de l'impact du réarrangement des attributs sur la performance des cartes.

— ***Ordre des données simulées***

Les données sont ordonnées selon le critère de dépendance et de séparation. L'objectif de cette section est de présenter l'ordre obtenu pour chaque critère.

La figure 5.5 montre les données simulées dans leur ordre d'origine, les données ordonnées avec le critère de dépendance soit, l'information mutuelle et la statistique de Pearson et les données ordonnées selon le critère de séparation. Pearson et le critère de séparation donnent exactement le même ordre. L'information mutuelle donne un ordre qui diffère au niveau des 3 premières variables, mais visualise plus ou moins les mêmes variables adjacentes. Ce

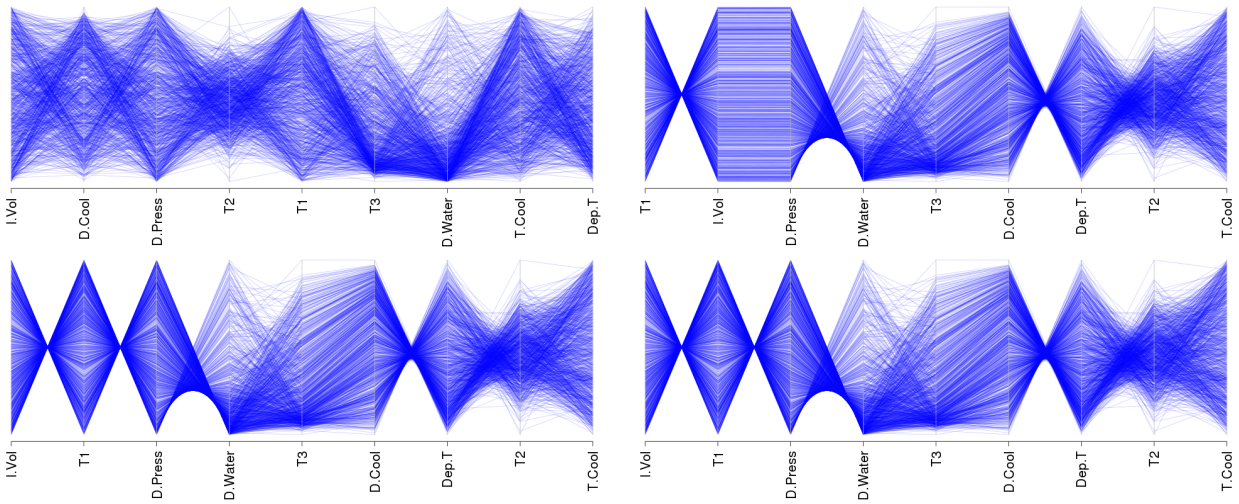


Figure 5.5 Figure illustrant les données dans l'ordre initial de simulation, les données ordonnées selon le critère de dépendance ; mesure d'information mutuelle et statistique de Pearson, et le critère de séparation.

qui est par contre très clair est que les données non ordonnées ne montrent pas de partern particulier, ne mettent pas en valeur les relations entre les variables et ressortent beaucoup de désordre. Dans le but d'évaluer l'impact de l'ordre des variables, nous gardons l'ordre trouvé par l'information mutuelle et celui trouvé par le critère de séparation. Pour expliquer la carte BOZ et analyser ses différentes composantes, nous utilisons uniquement l'ordre trouvé avec la statistique Pearson et le critère de séparation.

— ***Limites de la Best operating zone***

Avec les attributs ordonnés selon la mesure de séparation, les limites de la BOZ sont déterminées et finalement la BOZ est segmentée. Les paramètres de définition du modèle de cartes obtenus avec la méthode de validation croisée sont : La figure 5.6 présente un exemple des limites supérieures et inférieures de la BOZ des 9 attributs simulés. Une observation à l'extérieur de cette zone est classée comme un défaut. La BOZ est délimitée par des courbes et des lignes. Ces courbes et lignes traduisent en quelque sorte les limites des relations entre chaque paire de variables adjacentes. Elles sont particulièrement détectées suite au nouvel ordre des attributs. La relation entre $I.Vol$ et $T1$ est linéaire avec un coefficient négatif, ce qui est indiqué par 2 triangles. Un quotient reliant $D.Press$ et $D.water$ est représenté graphiquement par une courbe circulaire et un triangle. La relation linéaire à coefficient négatif est traduite par 2 lignes. 2 autres lignes correspondent à une relation de puissance d'ordre 4 entre $T3$ et $D.Cool$. Les 2 triangles qui apparaissent entre $Dep.T$ et $D.Cool$ signifient une racine carrée avec un coefficient négatif. Enfin, une fonction racine d'ordre 4 à coefficient négatif est illustrée par 2 lignes entre $Dep.T$ et $T2$. Les variables $T.Cool$ et $I.Vol$ et $D.Press$ et $T2$ sont 2

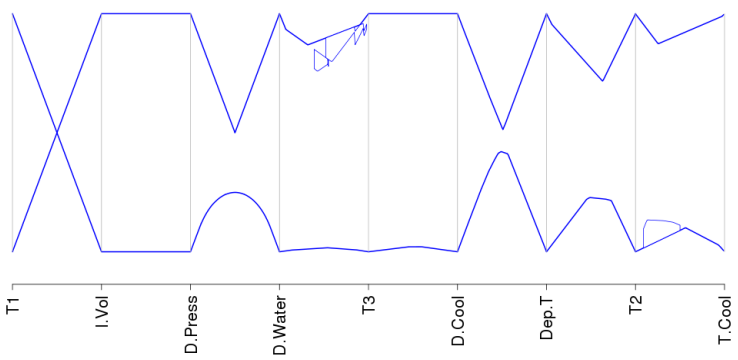


Figure 5.6 Figure illustrant les limites de BOZ.

par 2 indépendantes. Entre ces variables, des petites formes apparaissent. Ces formes correspondent à des zones vides dans les données historiques, des zones par lesquelles aucun point ne passe. Lorsque ces zones sont trop petites comme montré dans la figure 5.6, une hypothèse plausible est que ces zones vides sont du à un manque de données d'apprentissage. Cependant, le fait que ces zones n'apparaissent qu'entre les variables indépendantes, dans ce cas-ci, laisse cette hypothèse pas assez plausible. Une autre explication possible est que comme les variables sont indépendantes, les courbes entre celles-ci ne sont pas supposées prendre une forme bien définie. Les limites entre les variables indépendantes correspondent probablement aux limites de chacune des variables séparément. L'explication des formes observées sur les limites de la BOZ rejoint l'explication discutée dans la section 2.2.2.

Si les formes vides à l'intérieur des limites inférieures et supérieures sont assez larges, alors, 2 cas possibles peuvent avoir lieu :

— Soit il faut redéfinir les plans échantillonnage pour couvrir ces zones vides

— Soit les zones en question sont des zones de défauts

La redéfinition des plans d'expérience ne fait pas partie de cette thèse, elle nécessite la connaissance d'un expert du domaine du processus étudié. Ainsi, dans la manière dont nous définissons la BOZ, ces zones sont considérées comme des zones de défauts, si elles sont assez larges par rapport à l'échelle globale du graphique. "Assez large" est défini avec la méthode de validation croisée. Dans les sections suivantes, nous discutons la détection des défauts avec les cartes BOZ.

— ***Détection de dérives avec les limites de la BOZ***

L'analyse de la carte de contrôle BOZ montre que les limites de la BOZ permettent la détection de 2 catégories de défauts. La première catégorie est illustrée dans la figure 5.7. Cette figure représente des observations qui ne respectent pas les limites supérieures et inférieures

des variables. L'observation représentée en trait interrompu. dépasse la limite supérieure de la variable $D.Water$. Cependant, l'observation en trait continue dépasse la limite inférieure de la variable $T2$. Ainsi, la première catégorie de défauts représente les observations qui ne respectent pas les limites de chaque variable.

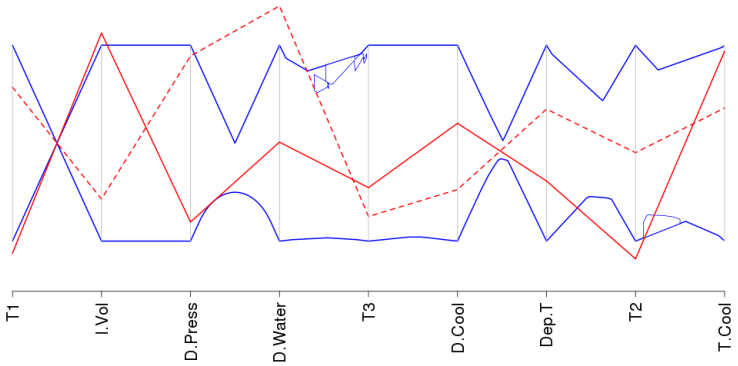


Figure 5.7 Figure représentant la première catégorie de défauts qui sont ceux qui dépassent la limite inférieure ou supérieure d'une ou de plusieurs variables.

La deuxième catégorie de défauts est illustrée dans la figure 5.8. Les points représentés respectent les limites supérieures et inférieures de de chaque variable, mais elles ne respectent pas limites correspondant aux relations entre certaines variables adjacentes. Dans la figure 5.8, l'observation en trait interrompu dépasse les limites correspondant à la relation entre $D.Press$ et $D.Water$ même si les limites supérieures et inférieures des deux variables sont respectées. Le point représenté en ligne continue dépasse ne respecte pas la forme de la relation entre $D.Cool$ et $Dep.T$ et la relation entre $Dep.T$ et $T2$.

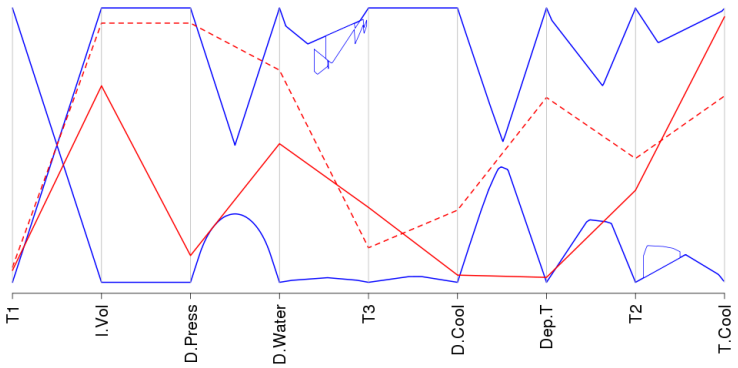


Figure 5.8 Figure illustrant le deuxième type de dérives soit les observations qui ne respectent pas les limites des relations entre les variables.

Ainsi, la deuxième catégorie de défauts détectés uniquement avec les limites de la BOZ

correspond aux observations qui ne respectent pas les limites des relations entre les variables. Les 2 catégories de défauts discutés sont détectées avec l'enveloppe de la BOZ. Toutefois, la BOZ est la zone où les observations ont une plus grande probabilité d'être fonctionnelles, mais ce n'est pas la zone sous contrôle (avec certitude). Cela signifie qu'une observation à l'intérieur de ces limites peut être un défaut. La segmentation est proposée dans l'objectif de réduire la probabilité de non-détection de défauts.

— ***Détection des défauts à l'aide des segments de fonctionnement***

Dans cette sous-section, nous expliquons l'importance de la segmentation dans la détection de défauts. Nous étudions, alors les défauts qui ne peuvent être détectés qu'avec les segments de fonctionnement. La figure 5.9 montre une même observation représentée avec les limites de la BOZ et en parallèle la même observation sur la carte de contrôle complète (limites de la BOZ et les segments).

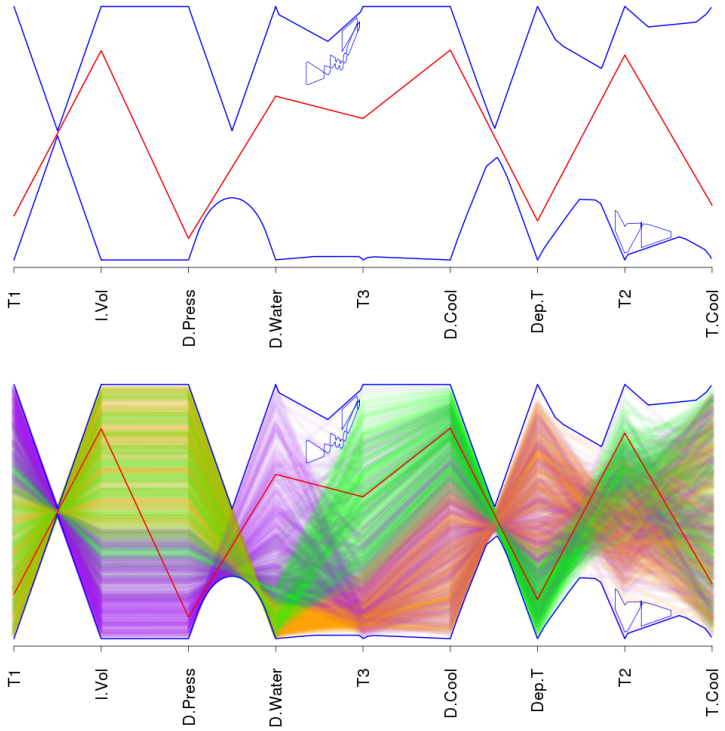


Figure 5.9 Figure illustrant les dérives détectables à l'aide des segments de fonctionnement.

Dans la figure de dessus, l'observation respecte les limites supérieures et inférieures de tous les attributs, ainsi que les limites des relations entre les attributs adjacents. Elle est à l'intérieur de la courbe enveloppe de la BOZ. Basée uniquement sur les limites de la BOZ, l'observation est classée comme fonctionnelle. Cependant, la figure de dessous montre que la même observation ne suit pas le modèle suivi par les autres observations fonctionnelles. Ceci est

détecté grâce aux segments de fonctionnement. L'observation n'appartient à aucun segment en particulier. Entre la variable $I.Vol$ et $D.Press$, l'observation passe du segment vert au segment violet. Par suite, avec la carte BOZ complète, l'observation est classée comme un défaut.

— ***Évaluation des cartes BOZ et de l'impact de l'ordre des attributs sur les cartes BOZ et Comparaison avec la carte d'Hotelling***

Nous présentons les résultats de classification des nouvelles observations avec la carte BOZ présentée dans le chapitre 4. La performance des cartes BOZ avec des attributs non-ordonnés avec des attributs ordonnés avec la mesure de séparation et des attributs ordonnés avec l'information mutuelle. Finalement, cette carte est comparée aux cartes $T^2$ d'Hotelling. Les cartes d'Hotelling sont générées comme expliqué dans le chapitre 2. La classification est faite automatiquement. L'évaluation continue avec la même base de données simulées.

Le tableau 5.2 représente les taux de classification correcte des différentes cartes. Le meilleur taux de classification est offert par les cartes BOZ développées en fonction des attributs ordonnées avec le critère de séparation (et la statistique Pearson) et avec l'information mutuelle. Les cartes avec des attributs ordonnées donnent un meilleur taux de classification que les autres cartes soient la carte d'Hotelling et la carte BOZ avec des attributs non-ordonnés. Ces cartes permettent de classer correctement les données dans 87.5% des cas comparativement à 81.36% lorsque les attributs ne sont pas ordonnés. La carte d'Hotelling donne un taux de classification correcte plus faible que toutes les cartes BOZ soit 79.99%. Les cartes d'Hotelling et les cartes BOZ avec attributs non réordonnés génèrent un pourcentage de fausses alarmes plus faibles avec les cartes BOZ avec attributs réordonnés, soit autour de 7% comparé à 9%. Les cartes BOZ avec données ordonnées permettent un taux de détection de défauts significativement meilleur que le taux des cartes d'Hotelling soient, respectivement, 84% par rapport à 67% et 70%. Ce qui peut expliquer le taux de fausses alarmes, assez faible, générées par les cartes d'Hotelling et les cartes BOZ sans ordre est la définition de limites de contrôle assez larges (relaxées). Ceci est confirmé par le taux de détection de défauts qui est également faible.

Les cartes BOZ avec données ordonnées offrent un résultat équilibré entre les données fonctionnelles et les défauts. Elles donnent un meilleur taux de classification global.

En conclusion, l'ordre des attributs a un impact sur le développement des cartes de contrôle. Le taux de classification correcte s'améliore d'une façon remarquable lorsque les attributs sont ordonnées. D'un autre côté, les cartes BOZ performent considérablement mieux que les cartes d'Hotelling.

— ***Expérience des utilisateurs***

Cette section décrit un test d'usage. Le but est d'évaluer la facilité d'utilisation de l'outil et

Tableau 5.2 Taux de classification correcte des différentes cartes étudiées.

| | | Taux de classification correcte | | |
|---|---|---|---|---|
| Type de carte | Ordre | global | des observations fonctionnelles | de défauts |
| Carte BOZ | ordre de départ | 81.36% | 92.72% | 70% |
| | Pearson et séparation | 87.5% | 90.45% | 84.54% |
| | Information mutuelle | 87.5% | 90.45% | 84.54% |
| Hotelling | | 79.77% | 92.27% | 67.27% |

sa compréhension par des utilisateurs externes à l'équipe de développement. Également, la capacité de diagnostic et de détection de défauts ont été évaluées. Les cartes BOZ ont été testées par 20 utilisateurs qui ne font pas partie de l'équipe de développement. Ces utilisateurs ne sont familiers ni avec les cartes de contrôle multidimensionnelles ni avec les coordonnées parallèles. Un nombre de 80 cartes est testé avec une nouvelle observation chaque fois. Chaque utilisateur a vérifié 4 cartes de contrôle BOZ et 4 cartes mono-dimensionnelles. Ensuite, il a répondu aux questions suivantes :

— Est ce que l'observation est fonctionnelle (détection) ?

— Sinon, quelle est la cause du défaut (diagnostic) ?

Le temps de détection et diagnostic a été chronométrée pour chacune des cartes. Par exemple, un utilisateur a vérifié la carte représentée sur la figure 5.10. Il a conclu que l'observation représente un défaut et son diagnostic disait que la limite entre $D.Cool$ et $Dep.T$ n'est pas respectée. La détection et le diagnostic ont pris 6 secondes. En moyenne, le taux de classifica-

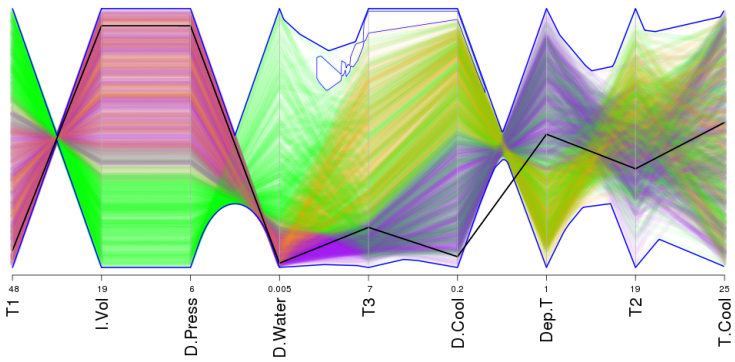


Figure 5.10 Figure illustrant le deuxième type de dérives soient les observations qui ne respectent pas les limites des relations entre les variables.

tion visuelle correcte est de 82%. Ce taux est plus faible que le taux de détection automatique. Ceci est principalement à cause du taux de fausses alarmes de 35%. Des observations fonctionnelles sont classées comme défauts. Dans la plupart des cas, les utilisateurs n'étaient pas sûrs du classement de l'observation. Ceci était principalement dû au désordre entre les variables $T.Cool$ et $I.Vol$ et entre $D.Press$ et $T2$. La structure entre ces variables n'est pas clair. Les utilisateurs avaient alors de la difficulté à juger si les observations entre ces variables doivent avoir un comportement bien précis ou pas, i.e, appartenir à un segment de fonctionnement en particulier. Donc, les utilisateurs ne pouvaient pas confirmer si une observation est fonctionnelle ou non. Par exemple, entre $D.Water$ et $T3$, l'utilisateur n'était pas sûr si l'observation sur la figure 5.11 est fonctionnelle ou non. Le désordre est dû à l'indépendance entre les attributs adjacents. Ceci aurait pu être réduit avec une meilleure explication de l'interprétation des segments entre les variables indépendantes ou avec des segments visualisés avec des polygones. Le taux de détection visuelle des défauts avec les cartes mono-dimensionnelles est de 40%. En effet, les cartes mono-dimensionnelles ne permettent pas de détecter les dérives liées aux relations entre les variables. Le diagnostic s'est fait correctement à de 100% avec les cartes BOZ. Le temps moyen mis pour répondre aux 2 questions posées est autour de 22 secondes pour les 2 cartes. Cette expérience montre que les cartes BOZ sont faciles à utiliser,

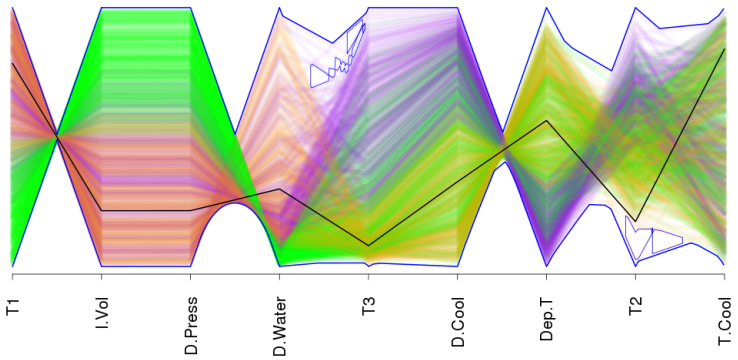


Figure 5.11 Figure illustrant le deuxième type de dérives soit les observations qui ne respectent pas les limites des relations entre les variables.

qu'elles facilitent le diagnostic des défauts détectés et elles n'exigent pas de connaissance particulière en mathématiques ou en maitrise des processus.
Cette section évalue les cartes BOZ à l'aide d'une base de données simulées. La section suivante continue l'évaluation avec une base de données réelle qui reflète un processus industriel réel.

### 5.2.2 Application à un cas réel : base de données de SPAM

Dans cette section, les cartes de contrôle basées sur les limites de la BOZ sont appliquées à la base de données SPAM et comparées aux cartes d'Hotelling.

— ***Description de la base de données***

Cette base de données ne reflète pas un processus de fabrication, mais le concept en reste proche. Les paramètres machines sont équivalents aux variables représentant le nombre d'apparitions de catégories de mots. Les classes non-spam, spam sont équivalentes aux classes d'observations fonctionnelles et défauts. La base de données de SPAM E-mail a été créée par Mark Hopkins et al. et offerte par George Forman . La base de données a été utilisée pour des objectifs de classification. Les spams inclus dans la base de données proviennent de postmaster et d'individus expéditeurs de spam. Les non-spams sont des e-mails professionnels ou personnels. Ces données sont utiles si nous désirons développer un filtre de courriels. Cette base de données contient 4601 observations dont 39.4% représentent des spams. Elle inclut 58 attributs dont 57 sont continues et 1 variable binaire qui indique la classe du courriel (spam ou non spam). Les attributs sont résumés dans le tableau 5.3.

La base de données est répartie en un sous-ensemble d'apprentissage de 1300 et un sous-ensemble de test de 1500 choisies aléatoirement. Les observations restantes n'ont pas été utilisées.

— ***Présentation des cartes de contrôle pour la base SPAM***

Pour développer les cartes de contrôle de la base de données de SPAM, les étapes expliquées dans le chapitre 4 sont suivies. Parmi les 57 variables, 50 sont choisies selon le critère de séparation ou de dépendance. L'ordre basé sur les données fonctionnelles avec le critère de dépendance permet d'éliminer les variables $V2$, $V19$, $V34$, $V46$, $V50$ et $V56$.

Cependant, l'ordre basé sur le critère de séparation élimine les variables suivantes $V4$, $V6$, $V17$, $V20$, $V48$, $V55$ et $V56$.

Pour en choisir un seul critère sur lequel la conception des cartes est basée, nous visualisons les données obtenues selon les 2 ordres proposés.

La figure 5.12 visualise les données SPAM ordonnées selon le critère de dépendance et de séparation. Les 2 critères d'arrangement donnent des résultats comparables au niveau de la visualisation des relations entre les variables, ainsi qu'au niveau de la séparation des données. L'ordre gardé est celui basé sur la séparation des données. La validation croisée en fonction des données d'apprentissage fonctionnelles et défaut a permis de définir les paramètres de la carte BOZ comme suit :

— les limites de la BOZ est défini à une distance $\epsilon_l = 0$ des minimums et maximums de

Tableau 5.3 Description des données de la base de SPAM.

| Attributs | Nombre et type | Description | Intervalle |
|---|---|---|---|
| $V1, .., V48$ | 48, réels | pourcentage de mots dans le courriel, un mot est un ensemble de caractère alphanumérique | $[0, 100]$ |
| $V49, .., V54$ | 6, réels | pourcentage de caractère CHAR dans le courriel, un mot est un ensemble de caractère alphanumérique | $[0, 100]$ |
| $V55$ | 1, réel | longueur moyenne de séquence continue lettres en majuscules | 1 |
| $V56$ | 1, entier | longueur de la plus longue séquence continue lettres en majuscules | $[1, ...]$ |
| $V57$ | 1, entier | nombre total de lettre majuscule | $[1, ...]$ |

la matrice de projection.

— Les limites des segments de fonctionnement sont définies à une distance $\epsilon_c = 0$ des minimum et maximums des segments de fonctionnement

— Les zones vides sont considérées lorsque leur largeur dépassent le seuil $s = 0.05$

— Les limites des zones vides sont définies en réduisant leurs largeurs d'une distance $e = 0.005$

Vu que les segments de données ne sont pas très bien séparables, nous choisissons de visualiser les segments avec les polygones colorés. Ainsi, la distinction entre les zones qui contiennent un seul segment et les zones qui contiennent plusieurs segments superposés est plus facile.

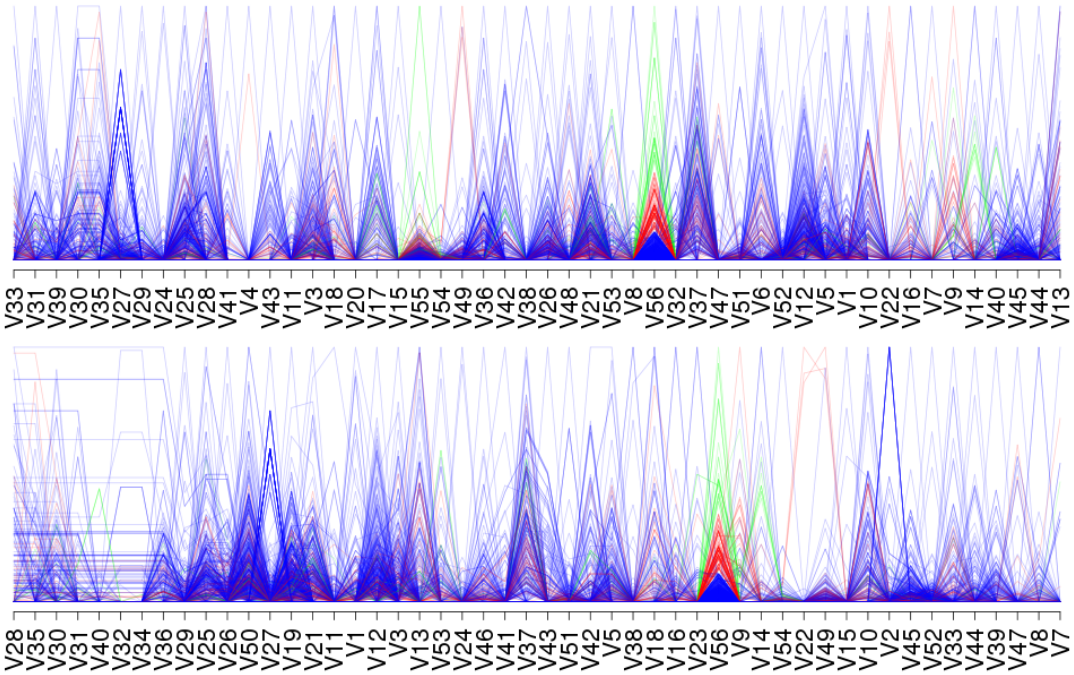


Figure 5.12 Données SPAM ordonnées avec le critère de séparation (figure en haut) et le critère de dépendance (figure de dessous).

Un segment est, ainsi, plus clair à limiter et à distinguer. Vu le nombre important de données de test, la classification des observations en données fonctionnelles et dérives a été faite automatiquement. Les cartes de contrôle sont utilisées pour diagnostiquer les défauts détectés (ou au moins certaines).

Un exemple de cartes développées est illustré dans la figure 5.13 Les limites de contrôle ne sont pas très lisses. De plus, beaucoup de zones vides apparaissent à l'intérieur de la BOZ. Au départ, une première analyse des cartes nous a poussés à penser que nous devons enlever les valeurs aberrantes des données d'apprentissage. Essayant de faire ceci, environ 20% des données allaient être supprimées. Ainsi, toutes les données sont gardées pour conception des cartes. Les cartes sont toujours celles avec des zones vides. Certaines zones de la BOZ incluent une seule valeur par observation. De plus avec ces données, nous avons obtenu des cartes avec des segments de fonctionnement plus ou moins superposés. En effet, il y a une région qui regroupe les 3 segments et quelques petites régions généralement proches des limites de la BOZ ou uniquement un segment apparaît. Par exemple entre les attributs $V_{35}$ et $V_{32}$ uniquement, pour les valeurs hautes, uniquement la couleur verte apparaît. Le fait que les limites ne soient pas lisses et que les segments ne soient pas trop distinguables est dû au fait que les attributs ne sont pas parfaitement dépendants et que les données ne sont pas séparables. Toutefois, ce fait n'empêche pas la classification des données. Pour classer

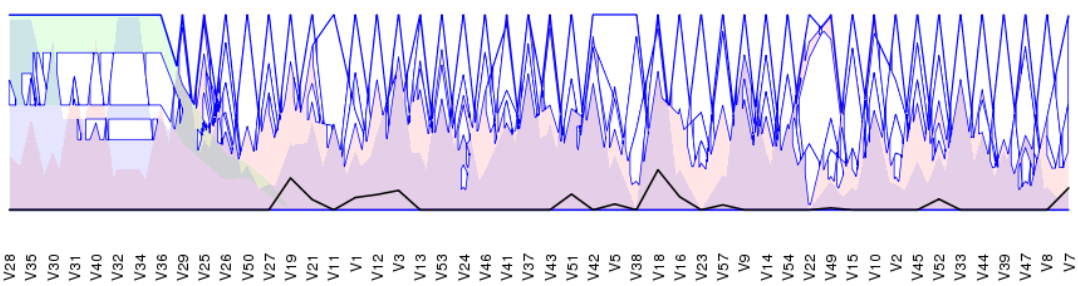


Figure 5.13 Exemple de carte de contrôle BOZ avec les données SPAM.

les nouvelles observations, nous nous fions aux quelques zones fonctionnelles séparables, aux limites de la BOZ et aux zones vides de la BOZ. Dans la section 5.2.2, les cartes BOZ des données SPAM sont évaluées et comparées aux cartes d'Hotelling.

— ***Évaluation de la carte de contrôle de SPAM***

La carte proposée appliquée à la base de données de SPAM est évaluée et comparée à la carte de Hotelling. Le taux de classification correcte, le taux de détection de défauts et le taux de fausses alarmes sont déterminées pour 2 types de cartes. Le tableau 5.4 résume ces taux de classification.

Tableau 5.4 Résultats de classification des données SPAM avec les cartes BOZ et la carte d'Hotelling.

| | Hotelling | outil proposé |
|---|---|---|
| Taux de classification correct | 69.3% | 76% |
| Taux de détection de défaut | 66.8% | 77.3% |
| Taux de fausses alarmes | 71.8% | 74.7% |

Le tableau 5.4 montre que les cartes BOZ donnent un meilleur taux de classification que les cartes d'Hotelling. En général, les cartes BOZ améliorent la classification des nouvelles observations de plus que 6%. De plus, elles génèrent 3% de moins de fausses alarmes et permet de détecter 77.3% des défauts comparés à 66.8% détectées avec les cartes d'Hotelling. Ce qui est, également, intéressant a remarquer est que les cartes BOZ garantissent un certain équilibre entre le taux de détection de défauts et le taux de classification des points fonctionnels. En

conclusion, les résultats donnés par la carte de contrôle BOZ sont meilleurs que ceux des cartes de contrôle de référence. Le diagnostic étant impossible avec les cartes de Hotelling, il est fait uniquement avec les cartes BOZ. Le diagnostic des dérives montre que les défauts sont principalement dus à des dépassements des limites supérieures ou à un passage dans les zones vides au niveau des variables $V_1$, $V_2$ et $V_{44}$.

## 5.3 Comparaison de la performance de la carte BOZ avec la carte d'Hotelling (ARL)

Le critère de performance choisi pour évaluer la carte de contrôle développée est le temps moyen de parcours ou encore Average Run Length en anglais (ARL). Le calcul de l'ARL pour la carte développée et la carte de Hotelling est faite à l'aide de la méthode de simulation Monte-Carlo comme expliquée dans la section 2.1.3. Les données sous contrôle sont simulées à l'aide de la loi normale. Les données sous contrôle ont pour moyenne $\mu = (0, 0, 0)$ et une variance $\mathbf{\Sigma}$. Les données utilisées pour estimer l'ARL sont aussi simulées à partir d'une loi normale avec une moyenne $\mu_1 = \mu + d$ ou $d$ est la distance de la moyenne sous contrôle et la même matrice de variance-covariance. L'ARL est évaluée pour différentes valeurs de $d$ soient $d \in [0, 0.5, 1, 1.5, 2, 2.5, 3, 4]$ et pour 3 valeurs de $\sigma = 0.2, 0.5, 0.8$. La variation des valeurs de $d$ a pour objectif d'évaluer la performance des cartes avec de petites et de grandes dérives par rapport à la moyenne. La variation de $\sigma$ vise à tester les cartes avec des variables corrélées, peu corrélées et indépendantes. Les valeurs de l'ARL pour tous les cas présentés sont résumées dans le tableau 5.5. L'évolution de l'ARL en fonction de la distance par rapport à la déviation de la moyenne est présentée par la figure 5.14.

Pour des données non corrélées soient $\sigma = 0.2$ et $\sigma = 0.5$, $ARL_0$ et $ARL_1$ pour les grandes dérives ne sont pas statistiquement différents entre les cartes d'Hotelling et les cartes BOZ. Pour les petites dérives, la carte BOZ donne de plus faibles ARLs. Lorsque les données sont corrélées ($\sigma = 0.8$), la carte BOZ est significativement meilleure en termes des $ARL_0$ et $ARL_1$. La carte d'Hotelling est performante car les données simulées suivent une loi symétrique, ce qui représente les meilleures conditions pour un bon fonctionnement de ce type de cartes.

## 5.4 Conclusion

En conclusion, les cartes BOZ sont aussi performantes que les cartes d'Hotelling de points de vue détection de défaut et en fausses alarmes. La carte est évaluée à l'aide de 2 bases de données. Le critère de performance, ARL, est évalué sur une base de données de taille réduite confirmant ainsi, que la carte BOZ a une performance comparable, voire meilleure

Tableau 5.5 Valeurs de L'ARL pour différentes covariances entre les variables et pour différentes distances de la moyenne.

| | $\sigma = 0.2$ | | $\sigma = 0.5$ | | $\sigma = 0.8$ | |
|---|---|---|---|---|---|---|
| Distance $d$ | Carte BOZ | Hotelling | Carte BOZ | Hotelling | Carte BOZ | Hotelling |
| $d = 0$ | 157.38 ±11.69 | **170.27** ±7.48 | 160.17 ±2.15 | **160.69** ±1.9 | **209.3** ±2.47 | 203.59 ±3.54 |
| $d = 0.5$ | **57.43** ±2.83 | 114.19 ±4.23 | **91.26** ±0.83 | 108.06 ±0.76 | **70.25** ±1.25 | 153.48 ±1.07 |
| $d = 1$ | **17.74** ±2.03 | 38.41 ±1.96 | **36.95** ±0.48 | 37.57 ±0.64 | **26.12** ±0.52 | 68.24 ±0.7 |
| $d = 1.5$ | 11.4 ±1.22 | **10.24** ±0.96 | **12.2** ±0.38 | 12.5 ±0.34 | **12.06** ±0.31 | 24.36 ±0.35 |
| $d = 2$ | 3.7 ±0.67 | **3.67** ±0.5 | 5.3 ±0.23 | **4.99** ±0.18 | **5** ±0.11 | 9.15 ±0.14 |
| $d = 2.5$ | **1.92** ±0.12 | 1.94 ±0.12 | 2.5 ±0.07 | 2.5 ±0.1 | **2.5** ±0.05 | 4.4 ±0.04 |
| $d = 3$ | 1.3 ±0.07 | **1.28** ±0.05 | 1.64 ±0.05 | **1.61** ±0.06 | **1.66** ±0.03 | 2.44 ±0.03 |
| $d = 4$ | 1.02 ±0.02 | **1.01** ±0.02 | 1.07 ±0.01 | **1.06** ±0.01 | **1.09** ±0.01 | 1.27 ±0.01 |

que la carte de référence avec un avantage assez importante qui le support de diagnostic. Cependant, comme discuté dans le chapitre 1, le développement des cartes BOZ nécessitent la disponibilité d'une base de données historique. Les cartes de contrôle densité sont présentées dans la section 5.5.

## 5.5 Évaluation des cartes de Contrôle de densité

L'objectif du développement des cartes de contrôle de densité est de faire face aux problèmes de collecte de données rencontrés notamment avec le partenaire industriel Teledyne Dalsa. Cette carte est développée avec un nombre limité de données fonctionnelles, uniquement. Le nombre de données n'est pas élevé. Cependant, le peu de données utilisées doit être assez représentatif de la population et doit provenir d'un processus assez contrôle et assez certain. Le test est conçu dans la vision d'évaluer la performance de la carte densité en fonction de la taille de la base de données historiques, entre autres, la courbe d'apprentissage en fonction

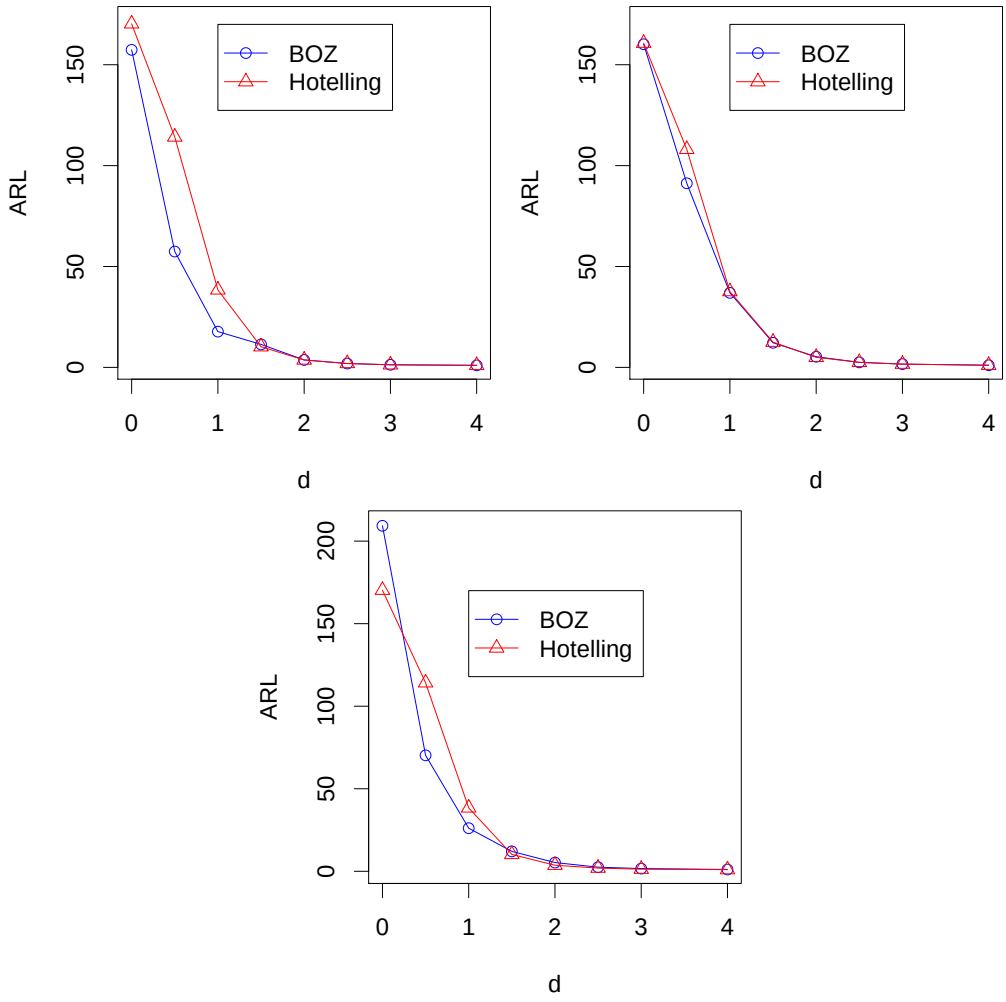


Figure 5.14 L'évolution de l'ARL en fonction de la distance dérivant par rapport à la moyenne sous contrôle. De gauche à droite, $\sigma$ varie de 0.2, 0.5 et 0.8.

de l'information disponible.

### 5.5.1 Base de données

Pour évaluer les cartes de contrôle densité, nous utilisons la même base de données utilisée dans la section 5.2. Or, l'objectif de cette section est de proposer des cartes qui fonctionnent avec un nombre de données limité. Ainsi, la base de données SPAM n'est pas utilisée en total. 540 parmi les 4040 observations sont sélectionnées aléatoirement pour constituer les données d'apprentissage. Pour refléter les processus très contrôlé, comme chez Teledyne Dalsa, ou la collecte de données de type défauts est plus complexe, la base de données historiques est constituée de 390 observations fonctionnelles et 140 défauts. Pour tester les cartes, 100 observations sont sélectionnées également aléatoirement. Les cartes basées sur ces données

sont présentées dans la section 5.5.2. Une comparaison avec les cartes de contrôle BOZ, les cartes densité, les cartes Hotelling et les réseaux de neurones est également présentée.

### 5.5.2 Présentation de la carte de contrôle densité

Les cartes de contrôle densité montrent un chemin de bon fonctionnement. Ce chemin est visualisé par les zones denses, soient les zones bleues et jaunes. Les observations qui passent à travers les zones rouges sont classées comme défauts telles que l'observation représentée sur la figure 5.15.
Lorsque le nombre d'observations est faible, les zones denses sont petites et ainsi la probabilité de classer une observation comme défaut est élevée. Lorsque le nombre d'observations historiques augmente, les zones denses deviennent plus larges augmentant, ainsi, la probabilité de classer un point comme fonctionnel. Il est important de trouver le nombre optimal de données à utiliser pour concevoir ce type de cartes.

### 5.5.3 Description du test

L'objectif du test est d'évaluer la courbe d'apprentissage des cartes densité en fonction de l'information disponible. Cette courbe est comparée à celle des réseaux de neurones, ainsi qu'à celle des cartes BOZ. L'information se modélise par le nombre d'observations de la base de données d'apprentissage (base de données historiques). L'apprentissage se traduit par le taux de classification correcte. Le but est de tester si les cartes sont performantes avec peu de données d'apprentissage et de déterminer ainsi la taille de la base de données historiques qui permet un apprentissage assez satisfaisant. Pour réaliser ce test, la base de données est répartie en 2 sous-ensembles. Le premier sous-ensemble est utilisé pour l'apprentissage. Ce sous-ensemble permet de développer les cartes. La deuxième base est la base de données de test. À partir de l'ensemble des données d'apprentissage, plusieurs sous-ensembles de données sont créés de façon à obtenir des bases de données de différentes tailles :

- $r$ observations sont aléatoirement choisies pour créer le premier sous ensemble.
- $r$ observations sont aléatoirement choisies parmi les observations restantes et ajoutées au premier sous ensemble pour constituer le deuxième sous ensemble et ainsi de suite

L'étape 2 est répétée jusqu'à ce que toutes les données de la base d'apprentissage soient incluses dans le sous-ensemble d'apprentissage. Pour chaque sous-ensemble de données les cartes densité sont développées à l'aide des données fonctionnelles. Le taux de classification correcte est déterminé, ainsi que le taux de détection de défaut et le taux de fausses alarmes.

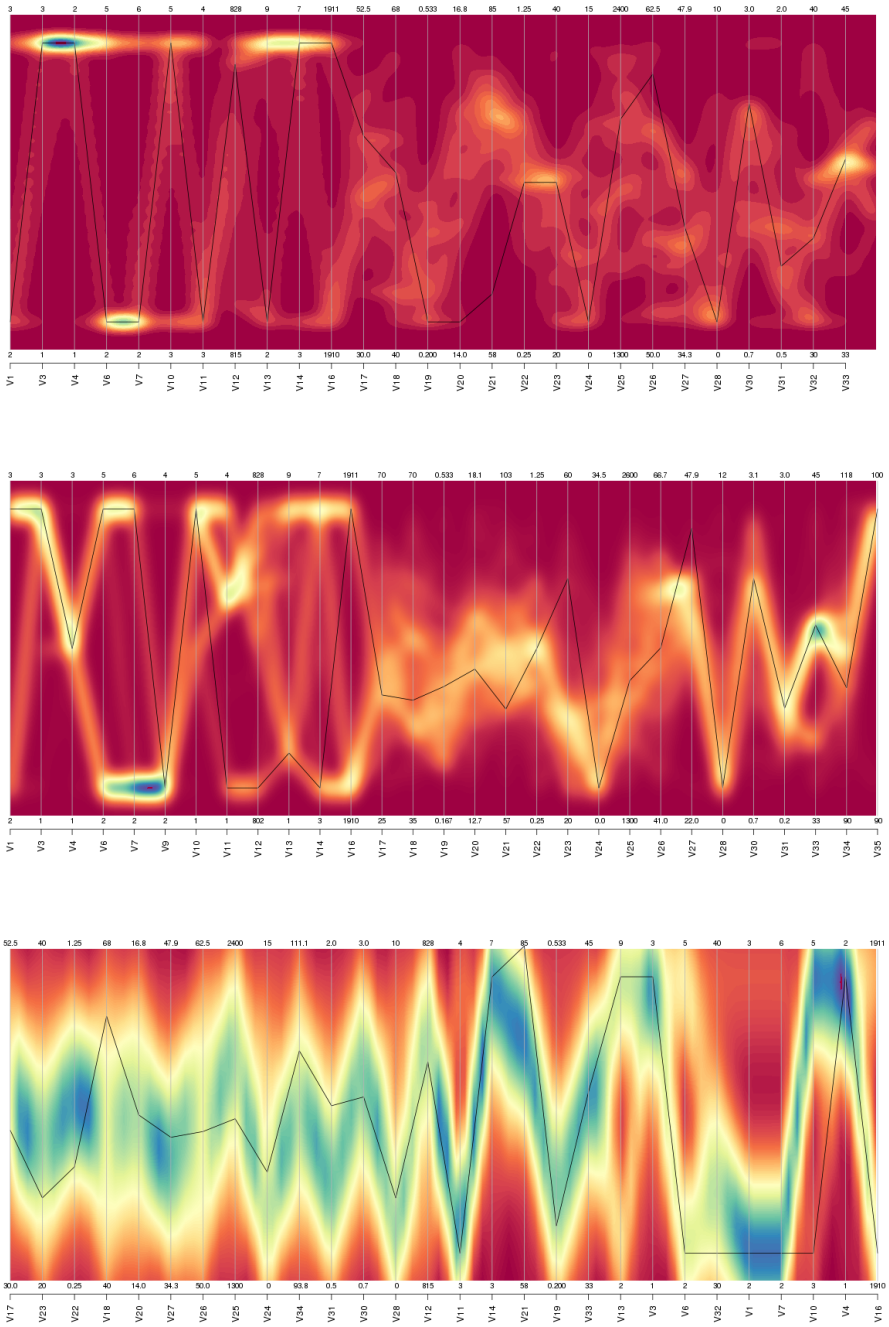

Figure 5.15 Figure illustrant les données dans l'ordre initial de simulation, les données ordonnées selon le critère de dépendance ; mesure d'information mutuelle et statistique de Pearson, et le critère de séparation.

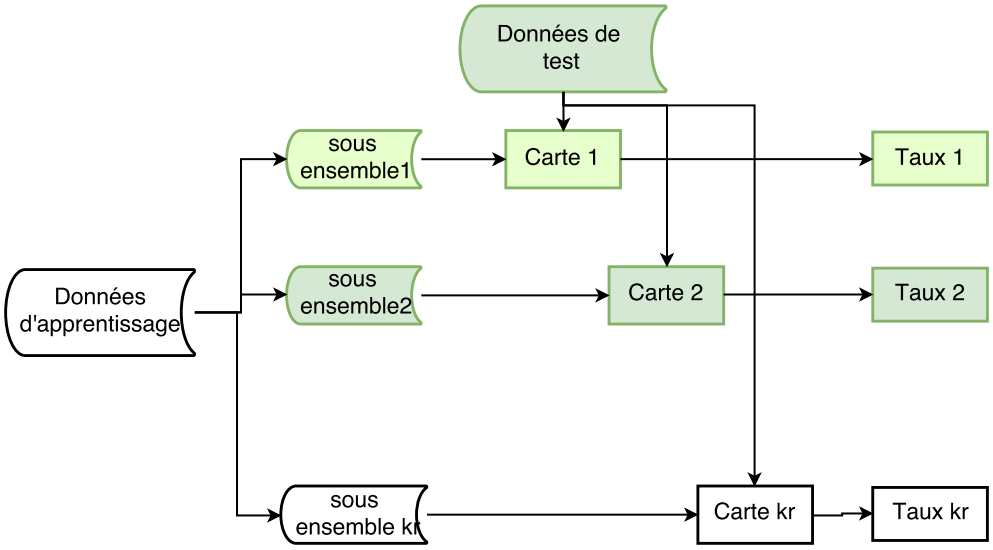


Figure 5.16 Méthode suivie pour tester les cartes de contrôle densité : courbe d'apprentissage.

### 5.5.4 Évaluation des cartes de contrôle densité

Dans cette section, la carte de contrôle densité est comparée aux réseaux de neurones. Les mêmes sous-ensembles d'apprentissage sont utilisés pour apprendre le modèle de réseaux de neurones. Pour chaque sous-ensemble d'apprentissage, le taux de classification correcte, le taux de détection de défauts et le taux de fausses alarmes sont évalués pour les cartes densité, les réseaux de neurones, les cartes BOZ et les cartes d'Hotelling. La figure 5.17 montre le taux de classification correcte des différents outils de contrôle. La figure 5.17 montre le taux de détection de défauts pour les 4 outils de contrôle. La figure 5.17 montre que la carte d'Hotelling donne un taux de classification correcte assez faible. Ceci peut être expliqué par 2 faits. Le premier est que les données historiques suivent une distribution non-symétrique. Le deuxième est que les défauts ne dérivent pas avec une grande distance en moyenne par rapport aux données fonctionnelles. Les cartes BOZ ne fonctionnent, également, pas très bien avec ces données. En effet, ce type de cartes est proposé pour le cas où les données historiques sont disponibles en quantité, ce qui n'est pas vérifié dans les conditions de ce test. Les comportements des cartes densité et des réseaux de neurones sont comparables. En effet, ces 2 outils fournissent le meilleur taux de classification. Le taux de classification des cartes densité augmentent jusqu'à atteindre le maximum et ensuite diminue. Pour un nombre d'observations

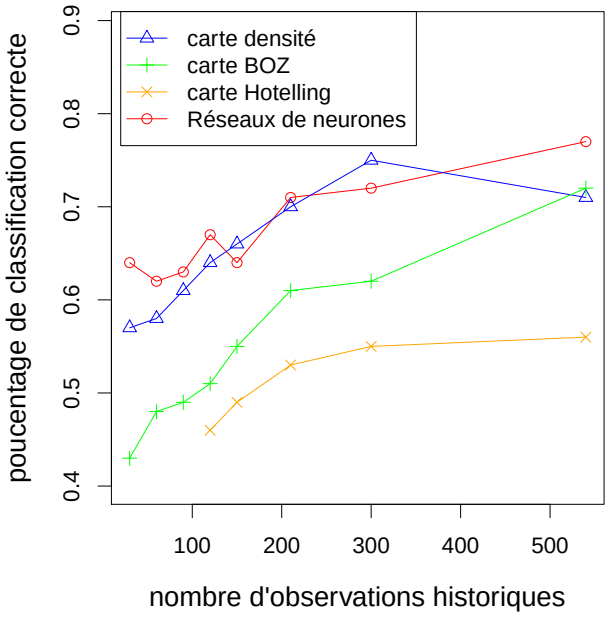


Figure 5.17 Courbe d'apprentissage des différents algorithmes testés, les cartes densité, BOZ et Hotelling et les réseaux de neurones

historiques inférieur à 120 observations, les réseaux de neurones fonctionnent mieux. Ensuite, le comportement devient identique pour les 2 outils jusqu'à 300 observations. Le meilleur taux atteint par les cartes densité soit d'environ 75% est atteint pour un nombre d'observations égal à 300 observations. Lorsque le nombre d'observations historiques dépasse 400 observations, les réseaux de neurones fonctionnent mieux. D'un autre côté, nous visualisons le taux de fausses alarmes générées par chacune des cartes (voir figure 5.18). De façon générale, le taux de fausses alarmes diminue lorsque le nombre d'observations historiques. Ceci veut dire que les modèles deviennent plus robustes avec de plus grandes bases historiques. Globalement, le taux de fausses alarmes est très élevé pour les cartes d'Hotelling et pour les cartes BOZ. Pour les cartes densité, le taux de fausses alarmes est élevé pour une taille de base historique inférieure à 100 et diminue exponentiellement à partir de 100 observations pour un taux plus faible que celui des réseaux de neurones. La zone dense ou encore la zone fonctionnelle est petite pour un nombre d'observations faible, ce qui explique le taux de fausses alarmes élevé. Par contre, lorsque le nombre d'observations historiques est élevé, la zone dense devient assez large de façon à ce que le taux de fausses alarmes devienne faible même comparé aux taux de classification incorrecte, soit 12% de fausses alarmes, pour 23% de classification incorrecte. En conclusion, les réseaux de neurones et les cartes densité fonctionnent d'une façon comparable. L'avantage des cartes BOZ par rapport aux réseaux de neurones est leur pouvoir de diagnostic et aussi leur performance avec un nombre limité de données historiques fonctionnelles qui est

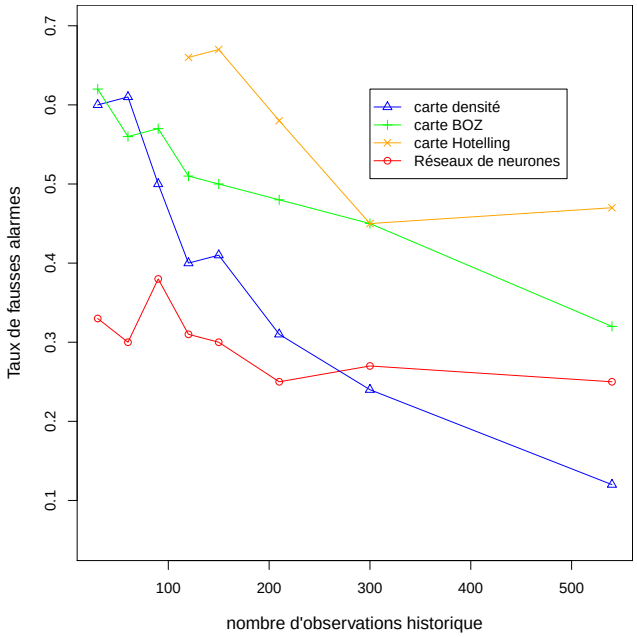


Figure 5.18 Évolution du taux de fausses alarmes des différents algorithmes testés, les cartes densité, BOZ et Hotelling et les réseaux de neurones en fonction du nombre d'observations de la base de données historiques.

le cas de plusieurs industries.

### 5.5.5 Lien avec les plans d'expérience

À partir de la figure qui illustre une carte de contrôle densité conçue avec un nombre de données égal à 20, les zones rouges sont très larges. Ainsi, ces zones peuvent soient représenter :

- des zones de défauts ;
- des zones de points fonctionnelles, mais non-explorées par les plans d'expérience définis lors de la collecte de données historiques.

Pour s'assurer de la signification des zones rouges, un expert du processus en question peut se baser sur les cartes pour étudier la possibilité de redéfinir les plans d'expérience de manière à explorer les zones non-couvertes.

### 5.5.6 Conclusion

Pour conclure, les cartes de contrôle développées performent aussi bien que les algorithmes reconnus dans la littérature. L'avantage primordial des cartes par rapport aux cartes de contrôle reconnues est leur pouvoir de diagnostic. Les cartes proposées dans cette thèse ont démontré une performance meilleure que la carte d'Hotelling et que les réseaux de neurones. Cepen-

dant, les cartes BOZ offrent un taux de classification correcte de 88%, mais elles nécessitent la disponibilité d'un nombre assez élevé de données historiques. Le critère de performance retenu pour évaluer cette carte est l'ARL. Les simulations démontrent que l'ARL de la carte BOZ est comparable, voire meilleur que celui des cartes d'Hotelling notamment l'ARL hors contrôle. La deuxième carte, étant la carte densité, est aussi performante. Elle fonctionne assez bien avec un nombre limité de données. Cependant, nous n'avons pas développé un algorithme de classification automatique. Ainsi, l'ARL n'a pas été évalué pour les cartes densité. Les cartes de densité ont été également testées avec les données du partenaire industriel. Le test a été réalisé dans l'objectif de démontrer l'applicabilité des cartes dans le contexte pour lequel elle a été conçue. Malheureusement, les données de test ne constituaient qu'une seule observation. Un tel test n'aurait pas pu démontrer la performance de la carte. Cependant, l'exemple de la carte du processus Dalsa est joint en annexe avec une brève description de la provenance des données.

# CHAPITRE 6 CONCLUSION

Cette thèse s'intègre dans le cadre d'un projet en partenariat avec Teledyne Dalsa et le C2MI. Elle est proposée pour soutenir les industriels dans leur processus de contrôle qualité et de détection de défauts.

## 6.1 Synthèse des travaux

Cette thèse propose un outil de maitrise de processus visuel et non paramétrique. Cet outil permet de concevoir 2 types de cartes de contrôle qui permettent la détection de défauts et soutiennent le diagnostic è l'aide de supports visuels basés sur les coordonnées parallèles. Le premier type de cartes est proposé pour le cas ou un l'ensemble de données historiques d'apprentissage est large et le deuxième est adapté au cas ou la collecte de données est plus compliquée résultant en un nombre limité d'observations. La conception des cartes de données passe par une première étape d'arrangement de variables. Cette étape vise optimiser la visualisation des données représentées en coordonnées parallèles. Elle aide è faciliter l'exploration visuelle des données, à ressortir les relations entre les variables et à améliorer la détection des segments de données. Le cadre présenté est générique, il s'adapte à l'objectif d'arrangement. Il se base sur le choix de 2 fonctions de probabilité qui définissent l'objectif d'arrangement, par exemple la dépendance entre les variables ou la séparation des données et sur le choix d'une fonction univariée qui définit la statistique, par exemple, la fonction $G(u) = ulog(u)$. Ce cadre peut s'adapter à plusieurs critères qui ne sont pas étudiés dans ce travail comme la minimisation des valeurs aberrantes. Appliqué pour réordonner les attributs dans l'objectif de séparer les données et dans l'objectif de mettre en valeur les dépendances, ce cadre général a démontré des résultats assez intéressants. Ce cadre a été développé pour les cartes de contrôle en particulier, mais peut être appliqué à différentes taches d'exploration visuelle de données.

Avec les données ordonnées, les cartes de contrôle sont développées. La première carte proposée que nous avons notée la carte BOZ, dans ce document, basée sur la caractérisation de la zone de bon fonctionnement ou encore la best operating zone par l'identification d'une sorte de courbe enveloppe qui délimite cette zone. Les points appartenant à cette zone ont une grande probabilité d'être fonctionnels. Les points à l'extérieur de cette zone sont classés comme défauts. Pour améliorer la précision de la BOZ, celle-ci est segmentée en de plus petits segments de fonctionnement. Ainsi, basée sur cette carte, si un point est à l'extérieur des limites de la BOZ ou qu'il n'appartient à aucun segment de fonctionnement, il est classé

comme défaut. Les cartes de contrôle BOZ, appliquées à une base de données et une base de données réelles, ont montré des résultats très intéressants par rapport aux cartes conventionnelles d'Hotelling. La détection de défaut pour ce type de cartes est supportée par un algorithme automatisé, ce qui a permis de simuler la longueur opérationnelle moyenne de cette carte qui est comparable à celle des cartes d'Hotelling, dans le cas ou les variables suivent une loi Normale.

Le deuxième type de carte est proposé pour le cas ou la collecte de données historiques est complexe. Cette carte est basée sur la projection des données en coordonnées parallèles dans un plan Cartésien bidimensionnel, le calcul de la fonction de densité par la méthode de noyaux et finalement la représentation de la fonction de densité obtenue. Les zones denses sont alors, représentées en bleu et les zones en rouge représentent les zones moins denses. La couleur des différentes zones passe du rouge au jaune au bleu lorsque la densité augmente. Un point est classé comme défaut s'il passe à travers une zone rouge. Ce deuxième type de cartes a, également, montré des résultats assez intéressants comparé aux cartes d'Hotelling et aux réseaux de neurones. Ainsi, les cartes développées montrent des taux de détection de défauts élevés et des taux de fausses alarmes faibles. Elles sont non paramétriques, tiennent en compte les relations entre les variables et ont un avantage très important qui est le support de diagnostic de défauts.

## 6.2 Limitations de la solution proposée

Malgré les avantages des algorithmes proposés, certaines limites sont à noter. Pour l'algorithme d'arrangement de variables, la solution proposée est une solution sous optimale (proche de la solution optimale), mais pas optimale. La solution optimale est longue à compiler et nécessite de la connexion entre R et CPLEX. La carte BOZ nécessite la disponibilité d'une large base de données historiques. Le développement de cette carte est basé sur une approche géométrique de projection de graphe en coordonnées parallèles plutôt que sur une approche statistique. Ainsi, l'ARL était évalué avec une simulation de Monte Carlo. L'évaluation théorique n'était pas faite. Pour les cartes densité, il y a un certain nombre d'observations ou elles sont optimales dépendamment de la nature des données. Le processus doit être assez bien contrôlé de telle façon que les classes des données historiques soient connues avec une bonne précision. La détection des défauts avec les cartes BOZ se fait visuellement. La détection n'est pas automatisée. Ceci empêche la détection en temps réel des défauts. De cette façon, les cartes BOZ sont destinées plutôt au diagnostic qu'à la détection. Le calcul de l'ARL était également impossible, même par simulation.

## 6.3 Améliorations futures

Le cadre proposé peut, également, inclure des critères d'arrangement qui tiennent en compte de la variable de sortie. La variable de sortie peut être par exemple la classe de données (fonctionnelle ou défaut). Ceci peut être en remplaçant le critère par un critère conditionnel par rapport à une variable de sortie, par exemple, la séparation des données conditionnellement à la classe de données. Un autre axe intéressant reste à explorer est celui de l'arrangement des données par rapport à sa contribution à la classe de sortie. Du coté des cartes de contrôle, les cartes de contrôle ne tiennent actuellement pas le temps en considération. Un axe à étudier serait la prise en compte du temps. Des pistes possibles sont le rajout d'un axe représentant le temps ou les cartes de contrôle en 3 dimensions. Un algorithme de détection automatique des défauts serait de grand intérêt pour les cartes densité. Il permettra une détection automatique, ainsi qu'une évaluation de l'ARL pour ces cartes. Également, l'implantation de ces cartes dans un processus réel demanderait l'ajustement et l'adaptation de l'algorithme au processus et aux données du processus.

# RÉFÉRENCES

Abouzahir, O., Gautier, R., and Gidel, T. (2003). Pilotage de l'amélioration des process par les coûts de non-qualité. *10ième Séminaire CONFERE, Belfort, France*, 3–4.

Albazzaz, H., and Wang, X. Z. (2006). Historical data analysis based on plots of independent and parallel coordinates and statistical control limits. *Journal of Process Control*, *16*(2), 103–114.

Albazzaz, H., and Wang, X. Z., and Marhoon, F. (2005). Multidimensional visualisation for process historical data analysis : a comparative study with multivariate statistical process control. *Journal of Process Control*, *15*(3), 285–294.

Ankerst, M., and Berchtold, S., and Keim, D. A. (1998). Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *Information Visualization, 1998. Proceedings. IEEE Symposium on*. IEEE, 52–60.

Bakir, S. T. (2001). Classification of distribution-free quality control charts. *Proceedings of the Annual Meeting of the American Statistical Association*.

Bakir, S. T. (2004). A distribution-free shewhart quality control chart based on signed-ranks. *Quality Engineering*, *16*(4), 613–623.

Bassetto, S., and Siadat, A. (2009). Operational methods for improving manufacturing control plans : case study in a semiconductor industry. *Journal of intelligent manufacturing*, *20*(1), 55–65.

Bect, P., and Simeu-Abazi, Z., and Maisonneuve, P-L. (2015). Identification of abnormal events by data monitoring : Application to complex systems. *Computers in Industry*, *68*, 78–88.

Bersimis, S., and Psarakis, S., and Panaretos, J. (2007). Multivariate statistical process control charts : an overview. *Quality and Reliability Engineering International*, *23*(5), 517–543.

Berthold, M. R., and Hall, L. O. (2003). Visualizing fuzzy points in parallel coordinates. *IEEE Transactions on Fuzzy Systems*, *11*(3), 369–374.

Bleakie, A., and Djurdjanovic, D. (2013). Feature extraction, condition monitoring, and fault modeling in semiconductor manufacturing systems. *Computers in Industry*, *64*(3), 203–213.

Boogaerts, T., and Tranchevent, L., and Pavlopoulos, G. A., and Aerts, J., and Vandewalle, J. (2012). Visualizing high dimensional datasets using parallel coordinates : Application to

gene prioritization. *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*. IEEE, 52–57.

Brooks, R., and Thorpe, R., and Wilson, J. (2004). A new method for defining and managing process alarms and for correcting process operation when an alarm occurs. *Journal of hazardous materials*, *115*(1), 169–174.

Chakraborti, S., and Van der Laan, P., and Bakir, S. T. (2001). Nonparametric control charts : an overview and some results. *Journal of Quality Technology*, *33*(3), 304.

Chen, H., and Li, H., and Fang, Y., and Chen, Y. (2016). Anisotropic parallel coordinates with adjustment based on distribution features. *Journal of Visualization*, *19*(2), 327–335.

Chen, J., and Liu, K.C. (2002). On-line batch process monitoring using dynamic pca and dynamic pls models. *Chemical Engineering Science*, *57*(1), 63–75.

Chen, K-Y., and Chen, L-S., and Chen, M-C., and Lee, C-L. (2011). Using svm based method for equipment fault detection in a thermal power plant. *Computers in industry*, *62*(1), 42–50.

Cortez, P., and Cerdeira, A., and Almeida, F., and Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.

Cressie, N., and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 440–464.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, *30*(3), 291–303.

Dasgupta, A., and Kosara, R. (2010). Pargnostics : Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1017–1026.

D'Ocagne, M. (1885). *Coordonnées parallèles & axiales : méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles*. Gauthier-Villars.

Dunia, R., and Edgar, T. F., and Nixon, M. (2013). Process monitoring using principal components in parallel coordinates. *AIChE Journal*, *59*(2), 445–456.

Duong, T. (2007). ks : Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of Statistical Software*, *21*(7), 1–16.

Ferdosi, B. J., and Roerdink, J. BTM (2011). Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Computer Graphics Forum*. Wiley Online Library, vol. 30, 1121–1130.

Fortin, J. (1990). *Québec. Le défi économique*. PUQ.

Freeman, M. F., and Tukey, J. W. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, 607–611.

Friedman, J., and Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics New York.

Fua, Y-H., and Ward, M. O., and Rundensteiner, E. A. (1999). Hierarchical parallel coordinates for exploration of large datasets. *Proceedings of the conference on Visualization'99 : celebrating ten years*. IEEE Computer Society Press, 43–50.

Gajjar, S. and Palazoglu, A. (2016). A data-driven multidimensional visualization technique for process fault detection and diagnosis. *Chemometrics and Intelligent Laboratory Systems*, *154*, 122–136.

Gang, T. T., and Yang, J., and Zhao, Y. (2013). Multivariate control chart based on the highest possibility region. *Journal of Applied Statistics*, *40*(8), 1673–1681.

Golub, T.R., and Slonim, D.K., and Tamayo, P., and Huard, C., and Gaasenbeek, M., and Mesirov, J.P., and Coller, H. and Loh, M.L., and Downing, J.R., and Caligiuri, M.A., and Bloomfield, C.D., and Lander, E.S. (1999). Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537.

Gonzaga, J. C. B., and Meleiro, L. A. C., and Kiang, C., and Maciel Filho, R. (2009). Ann-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Computers & Chemical Engineering*, *33*(1), 43–49.

Heinrich, J., and Weiskopf, D. (2009). Continuous parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1531–1538.

Heinrich, J., and Weiskopf, D. (2013). State of the art of parallel coordinates. *STAR Proceedings of Eurographics*, *2013*, 95–116.

Hunter, J. S. (1986). The exponentially weighted moving average. *J. Quality Technol.*, *18*(4), 203–210.

Hurley, C. B., and Oldford, R.W. (2011). Eulerian tour algorithms for data visualization and the pairviz package. *Computational Statistics*, *26*(4), 613–633.

Inselberg, A., and Dimsdale, B. (1990). Parallel coordinates : a tool for visualizing multidimensional geometry. *Proceedings of the 1st conference on Visualization'90*. IEEE Computer Society Press, 361–378.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering : a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

Jakubek, S. M., and Strasser, T. I. (2004). Artificial neural networks for fault detection in large-scale data acquisition systems. *Engineering Applications of Artificial Intelligence*, *17*(3), 233–248.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jiang, Q., and Yan, X., and Huang, B., (2016). Performance-driven distributed pca process monitoring based on fault-relevant variable selection and bayesian inference. *IEEE Transactions on Industrial Electronics*, *63*(1), 377–386.

Johansson, J., and Forsell, C. (2016). Evaluation of parallel coordinates : Overview, categorization and guidelines for future research. *Visualization and Computer Graphics, IEEE Transactions on*, *22*(1), 579–588.

Johansson, J., and Ljung, P., and Jern, M., and Cooper, M. (2005). Revealing structure within clustered parallel coordinates displays. *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, 125–132.

Johansson, S., and Johansson, J. (2009). Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 993–1000.

Kang, P., and Kim, D., and Lee, H-j., and Doh, S., and Cho, S. (2011). Virtual metrology for run-to-run control in semiconductor manufacturing. *Expert Systems with Applications*, *38*(3), 2508–2522.

Kourti, T., Nomikos, P., and MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls. *Journal of process control*, *5*(4), 277–284.

Lowry, C. A., and Montgomery, D. C. (1995). A review of multivariate control charts. *IIE transactions*, *27*(6), 800–810.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, *34*(1), 46–53.

Lu, L.F., and Huang, M.L., and Zhang, J. (2016). Two axes re-ordering methods in parallel coordinates plots. *Journal of Visual Languages & Computing*, *33*, 3–12.

Lynn, S. (2011). *Virtual metrology for plasma etch processes*. Thèse de doctorat, NATIONAL UNIVERSITY OF IRELAND, MAYNOOTH.

MacGregor, J. F., and Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, *3*(3), 403–414.

Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J., and Young, J. C. (1997). Assessment of multivariate process control techniques. *Journal of quality technology*, *29*(2), 140.

Montgomery, D. C., Torng, J. C. C., Cochran, J. K., and LAWRENCE, F. P. (1995). Statistically constrained economic design of the ewma control chart. *Journal of Quality Technology*, *27*(3), 250–256.

Nomikos, P., and MacGregor, J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, *40*(8), 1361–1375.

Palmas, G., and Bachynskyi, M., and Oulasvirta, A., and Seidel, H. P., and Weinkauf, T. (2014). An edge-bundling layout for interactive parallel coordinates. *Visualization Symposium (PacificVis), 2014 IEEE Pacific*. IEEE, 57–64.

Peng, W., and Ward, M. O., and Rundensteiner, E. A. (2004). Clutter reduction in multidimensional data visualization using dimension reordering. *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*. IEEE, 89–96.

Peterson, J. D. (2002). Clustering overview.

Precup, R-E., and Angelov, P., and Costa, B. S. J., and Sayed-Mouchaweh, M. (2015). An overview on fault diagnosis and nature-inspired optimal control of industrial process applications. *Computers in Industry*, *74*, 75–94.

Qiu, P. (2008). Distribution-free multivariate process control based on log-linear modeling. *IIE Transactions*, *40*(7), 664–677.

Qiu, P., and Hawkins, D., (2003). A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. *Journal of the Royal Statistical Society : Series D (The Statistician)*, *52*(2), 151–164.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*(3), 239–250.

Sharma, P., and Kaur, M. (2013). Classification in pattern recognition : A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(4), 298–306.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, vol. 26. CRC press.

Sun, R., and Tsung, F. (2003). A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, *41*(13), 2975–2989.

Umit, F., and Cigdem, A. (2001). Multivariate quality control : A historical perspective. *Yilditz Technical University*, 54–65.

Van Long, T., and Linsen, L. (2016). Efficient reordering of parallel coordinates and its application to multidimensional biological data visualization. *Visualization in Medicine and Life Sciences III*, Springer. 309–328.

Venkatasubramanian, V., and Rengaswamy, R., and Yin, K., and Kavuri, S. N. (2003). A review of process fault detection and diagnosis : Part i : Quantitative model-based methods. *Computers & chemical engineering*, *27*(3), 293–311.

Wang, R. C., and Edgar, T. F., and Baldea, M., and Nixon, M., and Wojsznis, W., and Dunia, R. (2015). Process fault detection using time-explicit kiviat diagrams. *AIChE Journal*, *61*(12), 4277–4293.

Weese, M., Martinez, W., Megahed, F. M., and Jones-Farmer, L. A. (2016). Statistical learning methods applied to process monitoring : An overview and perspective. *Journal of Quality Technology*, *48*(1), 4.

Woodall, W. H., and Driscoll, A. R. (2015). Some recent results on monitoring the rate of a rare event. *Frontiers in Statistical Quality Control 11*, Springer. 15–27.

Woodward, R. H., and Goldsmith, P. L. (1964). Cumulative sum tests : theory and practice. *ICI Monograph*, *3*.

Yu, J. and Xi, L. and Zhou, X. (2008). Intelligent monitoring and diagnosis of manufacturing processes using an integrated approach of kbann and ga. *Computers in Industry*, *59*(5), 489–501.

Zhou, H., and Yuan, X., and Qu, H., and Cui, W., and Chen, B. (2008). Visual clustering in parallel coordinates. *Computer Graphics Forum*. Wiley Online Library, vol. 27, 1047–1054.

Zou, C., and Qiu, P. (2012). Multivariate statistical process control using lasso. *Journal of the American Statistical Association*.